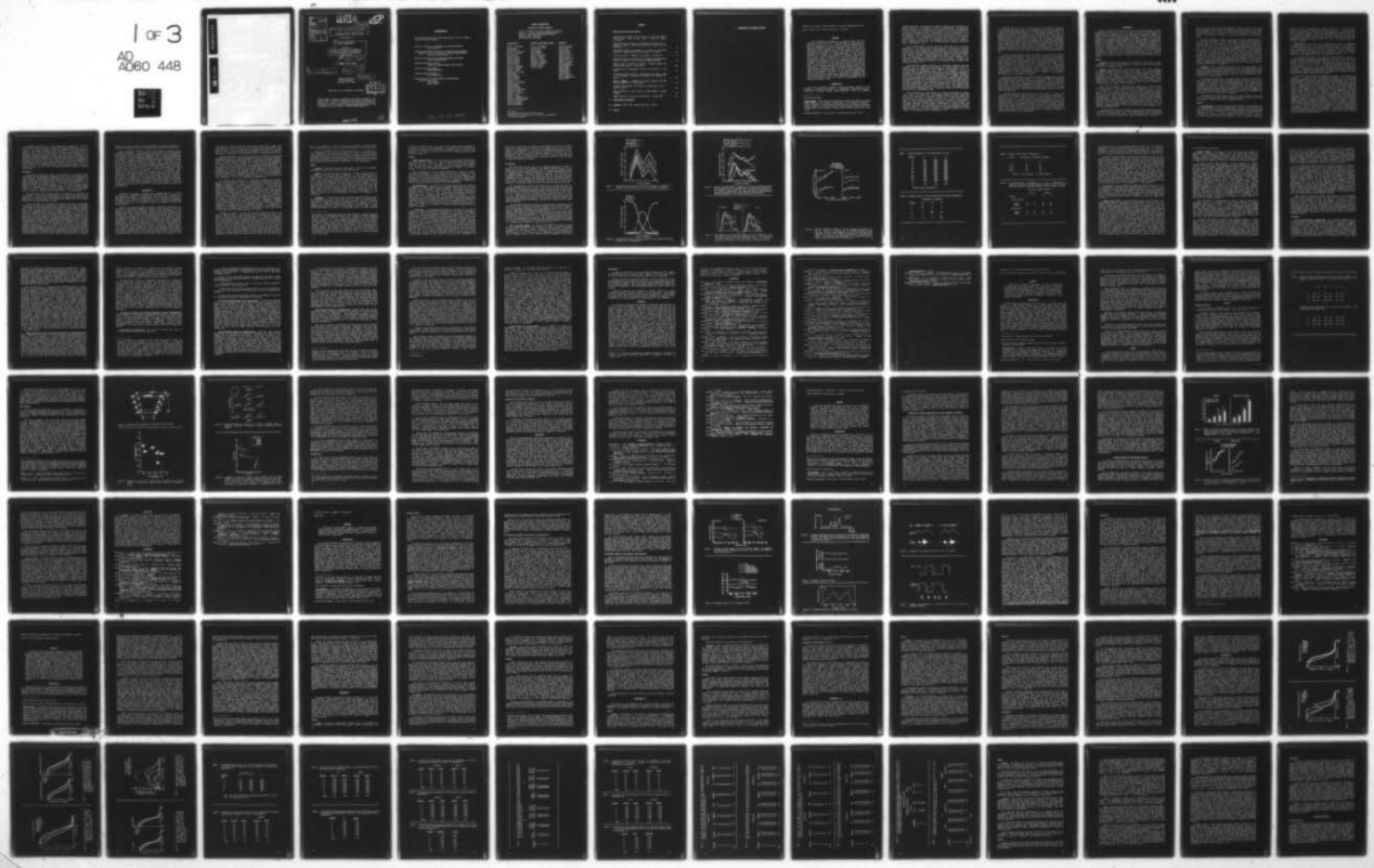
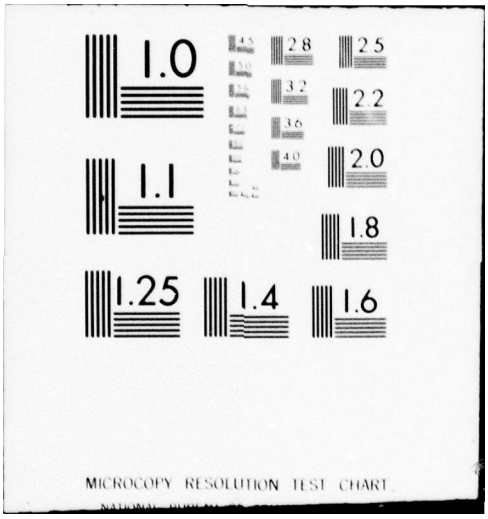
AD-A060 448 HASKINS LABS INC NEW HAVEN CONN F/G 5/7
SPEECH RESEARCH. A REPORT ON THE STATUS AND PROGRESS OF STUDIES--ETC(U)
JUN 78 A M LIBERMAN V101(134)P-342
UNCLASSIFIED SR-54(1978) NL

1 OF 3
AD A060 448




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
MICROCOPY RESOLUTION TEST CHART

LEVEL II

(14) SR-54 (1978)

(15) V101(134)P-342,
√PHS-HD-01994

Status Report on

(6) SPEECH RESEARCH.

A Report on
the Status and Progress of Studies on
the Nature of Speech, Instrumentation
for its Investigation, and Practical
Applications.

(9) Status rept.
1 Apr – 30 June 1978)

(10) Alvin M. Liberman

(11) 30 Jun 78

(12) 201 p.

Haskins Laboratories
270 Crown Street
New Haven, Conn. 06510

Distribution of this document is unlimited.

(This document contains no information not freely available to the general public. Haskins Laboratories distributes it primarily for library use. Copies are available from the National Technical Information Service or the ERIC Document Reproduction Service. See the Appendix for order numbers of previous Status Reports.)

406 643

1/B

ACKNOWLEDGMENTS

The research reported here was made possible in part by support from the following sources:

National Institute of Child Health and Human Development
Grant HD-01994

Assistant Chief Medical Director for Research and Development,
Research Center for Prosthetics, Veterans Administration
Contract V101(134)P-342

National Institutes of Child Health and Human Development
Contract N01-HD-1-2420

National Institutes of Health
Biomedical Research Support Grant RR-5596

National Science Foundation
Grant BNS76-82023

National Science Foundation
Grant MCS76-81034

National Institute of Neurological and Communicative
Disorders and Stroke
Grant NS13870
Grant NS13617

HASKINS LABORATORIES

Personnel in Speech Research

Alvin M. Liberman,* President and Research Director
Franklin S. Cooper,* Associate Research Director
Patrick W. Nye, Associate Research Director
Raymond C. Huey, Treasurer
Alice Dadourian, Secretary

Investigators

Arthur S. Abramson*
Thomas Baer
William Balch+
Fredericka Bell-Berti+
Gloria J. Borden*
Guy Carden*
Robert Crowder*
Steven B. Davis
Michael Dorman*
Donna Erickson*
William Ewan*
Carol A. Fowler*
Frances J. Freeman*
Jane H. Gaitenby
Thomas J. Gay*
Katherine S. Harris*
Alice Healy*
Hajime Hirose[1]
David Isenberg+
Leonard Katz*
Andrea G. Levitt
Isabelle Y. Liberman*
Leigh Lisker*
Virginia Mann+
Charles Marshall
Ignatius G. Mattingly*
Richard Pastore[2]
Lawrence J. Raphael*
Bruno H. Repp
Philip E. Rubin
Donald P. Shankweiler*
Michael Studdert-Kennedy*
Michael T. Turvey*
Robert Verbrugge*
Hirohide Yoshioka[1]

Technical and Support Staff

Eric L. Andreasson
Elizabeth P. Clark
Donald Hailey
Terry Halwes
Elly Knight*
Sabina D. Koroluk
Agnes M. McKeon
Nancy R. O'Brien
William P. Scully
Richard S. Sharkany
Leonard Szubowicz
Edward R. Wiley
David Zeichner

Students*

Linda D'Antonio
David Dechovitz
Laurel Dent
Laurie Feldman
Hollis Fitch
Carole E. Gelfer
Robb Gilford
Janette Henderson
Robert Katz
Morey J. Kitzman
Roland Mandler
Karen Marcarelli
Leonard Mark
Nancy McGarr
Georgia Nigro
Brad Rakerd
Abigail Reilly
Arnold Shapiro
Emily Tobey-Cullen
Betty Tuller

---

*Part-time
[1]Visiting from University of Tokyo, Japan
[2]Visiting from State University of New York at Binghamton
+NIH Research Fellows

CONTENTS

I. MANUSCRIPTS AND EXTENDED REPORTS

Categories and Context in the Perception of Isolated Steady-State Vowels

Bruno H. Repp, Alice F. Healy[+], and Robert G. Crowder[+]

ABSTRACT

The noncategorical perception of isolated vowels has been attributed to the availability of auditory memory in discrimination. Using vowels from an /i/-/I/-/ɛ/ continuum in an AX (same-different) task and comparing the results with predictions derived from a separate identification test, we demonstrate that vowels are perceived more nearly categorically if auditory memory is degraded by extending the interstimulus interval (ISI) and/or filling it with irrelevant vowel sounds. In a second experiment, we use a similar paradigm but, in addition to presenting a separate identification test, elicit labeling responses to the AX pairs used in the discrimination task. We find that AX labeling responses predict discrimination performance quite well, regardless of whether auditory memory is available or not, whereas the predictions from the separate identification test are more poorly matched by the obtained data. The AX labeling responses show large contrast effects (both proactive and retroactive) that are greatly reduced when there is interference with auditory memory. We conclude from the presence of these context effects that vowels are not perceived categorically (that is, absolutely). However, it seems that by properly taking these context effects into account, discrimination performance can be quite accurately predicted from labeling data, suggesting that vowel discrimination, just like consonant discrimination, may be mediated by phonetic labels.

INTRODUCTION

One of the best-known findings of speech perception research is the phenomenon of categorical perception. Its experimental demonstration requires a continuum of synthetic speech sounds spanning at least two different

---

[+]Also Yale University.

Acknowledgment: This research was supported by NICHD Grant HD01994 and BRSG Grant RR05596 to the Haskins Laboratories. We are greatly indebted to Virginia Walters for conducting the experiments and tabulating the results. Helpful comments on earlier versions of this manuscript were obtained from William Estes, Alvin Liberman, Virginia Mann, Richard Pastore and David Pisoni.

[HASKINS LABORATORIES: Status Report on Speech Research SR-54 (1978)]

phonemic categories. If listeners are asked to identify and discriminate the stimuli from such a continuum, two typical findings emerge: the perceptual boundary between the two categories is relatively abrupt, and discrimination of stimuli drawn from within the same category is near chance, while it is good across the phoneme boundary. Perception is said to be categorical if the discrimination results can be predicted from the identification results, under the assumption that discrimination is based exclusively on the phonetic category labels. This pattern of results has been typical for a number of speech sounds, particularly the stop consonants in initial position (Liberman, Harris, Hoffman and Griffith, 1957; Studdert-Kennedy, Liberman, Harris and Cooper, 1970; Pisoni, 1971).

Of all speech sounds, isolated vowels are least likely to be perceived categorically. Not only are the category boundaries less distinct on a vowel continuum, but, more significantly, discrimination of acoustically different stimuli from the same category is usually much better than chance and exceeds the predictions derived from labeling data (Fry, Abramson, Eimas and Liberman, 1962; Stevens, Liberman, Studdert-Kennedy and Öhman, 1969; Pisoni, 1971). Thus, the perception of isolated vowels has been called "continuous," in contrast to the categorical perception of stop consonants. However, even vowels often show better discrimination across category boundaries than within categories (Pisoni, 1971).

The distinction between categorical and continuous perception has been attributed to the differential availability of auditory memory traces for different kinds of stimuli (Darwin and Baddeley, 1974; Pisoni, 1971, 1973, 1975; Fujisaki and Kawashima, 1969, 1970). The assumption is that an accessible auditory memory representation facilitates continuous perception by providing an alternate basis for discrimination beyond phonemic categories. We will refer to this view as the dual-coding model. It assumes that speech sounds are discriminated by comparing both auditory and phonetic memory codes. The distinctive cues for stop consonants are of very brief duration and are followed (or preceded, if in final position) by a vowel that might interfere with the already fragile auditory memory representation of the relevant cues. Isolated steady-state vowels, on the other hand, are of much longer duration and contain distinctive information from onset to offset. Consequently, their auditory memory representation will be much more robust and can be utilized more easily in a discrimination task.

This explanation has found support in several experiments by Crowder (1971, 1973a). He showed that three standard phenomena of verbal short-term memory--the recency effect, the suffix effect, and the modality effect--are all obtained for lists of syllables differing only in their vowels, but are absent in lists of syllables differing only in their initial stop consonants. Since all three effects mentioned are assumed to reflect the existence of a relatively unanalyzed auditory memory (precategorical acoustic store), the conclusion was that stop consonants do not leave any significant auditory trace but vowels do. The relative strength of the auditory memory trace for a stimulus seems to be a direct function of its acoustic similarity to other stimuli to be remembered (Darwin and Baddeley, 1974).

Investigations of the role of auditory memory in categorical perception have taken two approaches. One line of research tries to make the perception of stop consonants less categorical by inducing listeners to make better use of their weak auditory memory traces. The other approach attempts to make the perception of vowels more categorical by interfering with their auditory memory representations, so that the listeners have to rely increasingly on category labels in discriminating the stimuli. The first type of experiment involves listener training and the use of sensitive discrimination paradigms; it has yielded some positive results suggesting that, under favorable conditions, listeners can make effective use of their auditory memory representations of stop consonants (Carney, Widin, and Viemeister, 1977; Ganong, 1977; Pisoni and Lazarus, 1974; Sachs and Grant, 1976; Samuel, 1977). The other approach--that of making the perception of vowels more categorical--is primarily due to Pisoni (1971, 1973, 1975), whose work provides the background for ours.

Prior to Pisoni's studies, there was already some evidence that vowels are perceived more categorically when they occur in phonetic (word) context (Stevens, 1968; Sachs, 1969). Thus, one way to decrease the strength of auditory memory is to change the structure of the stimuli themselves. Pisoni (1971, 1973), Fujisaki and Kawashima (1969, 1970), and Sachs (1969) took a related approach by decreasing the duration of isolated vowel stimuli. This made perception more categorical, but not completely so. Corresponding reductions in the stimulus suffix effect for shortened vowels were reported by Crowder (1973b) and Hall and Blumstein (1977). However, even very short vowels apparently permit distinguishable auditory traces to be established; discrimination is usually better than would be expected if only phonetic labels were used to discriminate the stimuli.

An alternative procedure is to leave the stimuli unchanged and to attempt to tamper directly with auditory memory. There are two methods that have been used to degrade auditory memory, decay and interference. The first technique was used by Pisoni (1971, 1973) and, more recently, by Cutting, Rosner, and Foard (1976). These authors systematically increased the interval between the vowel stimuli in an AX (same-different) discrimination task from 0 to 2 sec. The result was a decrease in performance that was taken to mean that auditory memory decayed over time. Whether this decay was complete after 2 sec is not clear from their data.

The interference technique was employed by Pisoni (1975). He used an ABX discrimination paradigm in which the "X" vowel was immediately preceded or followed by one of four irrelevant signals: a noise burst, a pure tone, a dissimilar vowel, or a similar vowel. Performance decreased in all conditions, but more so when the interfering stimulus followed the "X" vowel than when it preceded it. An acoustically similar vowel seemed to produce the most interference. These interference effects are similar to the phonetic context effects of Stevens (1968) and Sachs (1969), except for the fact the coarticulation with natural phonetic context modifies the acoustic properties of vowels, while unrelated interfering stimuli do not. We decided to employ both decay and interference in our experiments.

EXPERIMENT I

Pisoni's results are consistent with a role for auditory memory in vowel perception, but since Pisoni did not attempt to predict discrimination from identification performance, we do not know whether procedures designed to eliminate auditory memory would produce completely categorical perception for vowels. This is the hypothesis that we wished to test in our first experiment. To manipulate the availability of auditory memory, we varied the amount of time elapsing between members of AX (same-different) discrimination pairs and, orthogonally, whether there was an interpolated speech sound or not during this interval. To measure the degree of categorical perception, we relied on a comparison of identification and discrimination performance: if accuracy of AX discrimination can be predicted from phonetic labeling (identification), provided that both are better than chance, we may conclude that perception is categorical. The interesting possibility is that, although discrimination shows a surplus over identification when auditory memory is present, vowel perception will be categorical when auditory memory has been removed.

Method

Subjects. Sixteen college-age adults volunteered to participate as paid volunteers. All were native speakers of English and had little previous experience with synthetic speech.

Stimuli. The stimuli were modeled after Pisoni's (1971) vowel continuum. The formant frequencies given in Pisoni (1971, Table 2, p. 12) were realized as closely as possible (within a few Hz) on the OVE IIIc synthesizer at Haskins Laboratories. The complete set included 13 stimuli whose first formant increased, and whose second and third formants decreased in frequency in approximately equal logarithmic steps from stimulus 1 to stimulus 13. These frequencies are shown in Table 1. The fourth and fifth formants were hardware-fixed. All stimuli were 240 msec in duration and had a fundamental frequency that fell linearly from 125 to 80 Hz.

From these 13 stimuli, three pairs of vowels were selected which, according to Pisoni's data, were identified predominantly as /i/, /I/, and /ɛ/, respectively. In the notation of Table 1, they were stimuli 1 and 3 (/i/), 6 and 8 (/I/), and 11 and 13 (/ɛ/). Note that the acoustic distance between the vowels was greater between categories (three steps) than within (two steps); this was a deliberate attempt to avoid the fairly broad category boundary regions evident in Pisoni's data and to maximize between-category discriminability. Within-category discriminability of the stimuli selected had been about 80 percent correct in the ABX test used by Pisoni (1971). An additional vowellike sound was constructed by combining the first formant of stimulus 1 with the higher formants of stimulus 13 (see Table 1). This stimulus sounded approximately like the vowel /y/ (as in Swedish fyra) and was used for interference only.

Six experimental tapes were recorded using the Haskins Laboratories Pulse Code Modulation (PCM) System. Two were identification tapes; the remaining four were AX discrimination tapes. One identification tape contained 60 vowels (the six stimuli repeated ten times) in random order, with ISIs of 4

sec. The second identification tape contained the same 60 vowels in the same random sequence, but each vowel was preceded by the irrelevant /y/ sound. The interval between the /y/ and the following vowel was 120 msec; that between the vowel and the next /y/ was 4 sec. The /y/ precursor was included as a control to see whether it affected in any way the labeling of the following vowel.

The four discrimination tapes all contained the same random sequence of 80 vowel pairs consisting of five replications of sixteen different combinations of the six basic stimuli. The sixteen combinations included six identical pairs (1-1, 3-3, 6-6, 8-8, 11-11, 13-13), six within-category pairs (1-3, 3-1, 6-8, 8-6, 11-13, 13-11) and four between-category pairs (3-6, 6-3, 8-11, 11-8). The four tapes differed in the nature of the interval between the two vowels in a pair. In the "short unfilled" condition, it was 480 msec of silence. In the "long unfilled" condition, it was 1,920 msec of silence. In the "short filled" condition, the /y/ sound (240 msec in duration), preceded and followed by 120 msec of silence, intervened between the two vowels. In the "long filled" condition, five repetitions of the /y/ sound intervened; they were preceded, separated, and followed by 120 msec of silence. Thus, the temporal separation between the vowels in a pair was the same in corresponding filled and unfilled conditions. The interval between successive pairs was 4 sec throughout.

Procedure. The 16 subjects were divided into two equal groups. One group received the two identification tests prior to the discrimination tests; the other group was assigned the reverse order. All subjects listened to the regular identification series before the one with /y/ preceding each vowel. The sequence of the four discrimination conditions was counterbalanced across subjects in four Latin squares.

In the identification task, the answer sheets listed the words "beet," "bit," and "bet" for each trial. The subjects were instructed to circle the word whose vowel resembled most the stimulus presented. The /y/-sound was to be ignored, if present. In the discrimination tasks, the response sheet contained the letters "s" (same) and "d" (different) for each trial, and the subjects were instructed to circle the appropriate letter for each vowel pair. It was emphasized to respond "same" only when the two vowels were exactly the same. The different conditions were explained and announced in advance. Any occurrences of the /y/ sound were to be ignored.

The subjects were run in small groups in a single session of about one hour. The tapes were played back on a Sony TC-630 tape recorder with its own loudspeakers. Intensity was set at a comfortable level.

Results

Identification. The identification results, averaged across subjects and the two identification tests, are summarized in Table 2. The results of the two identification tests were combined, since an analysis of variance showed that the irrelevant /y/-precursor did not significantly affect identification performance, $F(1,14) < 1$. Table 2 shows that stimuli in the /i/ and /ɛ/ categories were identified fairly consistently (89 percent correct or better), but many confusions occurred with stimuli in the /I/ category, especially

stimulus 8. This is in agreement with Pisoni's (1971) data: the /I/ category is the least stable of the three, probably because the relatively long stimulus durations employed here were least appropriate for this category which, in natural speech, is associated with shorter durations than /i/ and /ɛ/ (Peterson and Leniste, 1960). The statistical analysis indicated that confusions were somewhat more frequent when the identification tests were presented at the end of a session, $F(1,14) = 7.3$, $p = .017$; this may have been a result of fatigue.

Discrimination. The results of the discrimination tests are summarized in Figure 1. For each of the four experimental conditions, percentages of correct responses are shown as a function of stimulus pair. Each data point is plotted halfway between the locations of the two stimuli to be discriminated and represents the average of four percentages: those of "different" responses to the two stimulus orders of the given pair, and those of "same" responses to each member of the pair when paired with itself.

It is evident from Figure 1 that both manipulations of the ISI (delay and filling) affected discrimination performance. The subjects made more errors when the interval was long than when it was short, $F(1,14) = 56.4$, $p < .0001$, and when the interval was filled with irrelevant vowel sounds than when it was unfilled, $F(1,14) = 40.0$, $p < .0001$. The interaction of these two factors was not significant, $F(1,14) < 1$, nor was there any significant interaction of these two effects with vowel pairs, as is confirmed by the parallel functions in Figure 1. Newman-Keuls tests between individual conditions confirmed that both delay by itself and the presence of an interpolated stimulus by itself significantly reduced discrimination performance.

As expected, discrimination performance was poorest in the long filled condition. In order to find out whether the scores in this condition approached those expected under the categorical perception model, we predicted the percentages of correct responses in the discrimination test from the identification responses, under the assumption that discrimination is based solely on phonetic labels (Pollack and Pisoni, 1971). (A second assumption, often not stated explicitly, is that the labeling probabilities are the same in the identification and discrimination tasks; we will have reason to question this assumption later in this paper.) These predictions, averaged across subjects, are indicated in Figure 1 by the dashed function at the bottom. They are quite close to the scores obtained in the long filled condition, particularly for the first three stimulus pairs; the discrimination performance for the last two vowel pairs is somewhat better than predicted. Separate Chi-square tests were performed on the data from each subject, summing observed and expected frequencies of correct responses and errors across stimulus pairs, thereby enabling tests with only one degree of freedom. When comparing expected scores to those obtained in the long filled condition, 11 of the 16 subjects exceeded the expectations, but the difference was significant at the .05 level in only two cases. In addition, another subject performed significantly poorer than expected. Thus, overall performance in the long filled condition was not significantly different from that predicted by the categorical-perception model.

A separate analysis of "hits" ("different" responses to pairs of nonidentical stimuli) revealed an unexpected stimulus order effect, $F(1,14) = 15.3$, $p < .01$, which is shown in Table 3, averaged across conditions and subjects. This effect interacted with position on the continuum, $F(4,56) = 10.4$, $p < .0001$. In four stimulus pairs, the subjects gave substantially more "different" responses when the stimulus with the higher position on the continuum preceded the stimulus with the lower position, but the effect was reversed for the last stimulus pair: there were more "different" responses to the order 11-13 than to 13-11. The stimulus order effect was somewhat more pronounced when the ISI was unfilled than when it was filled, leading to a significant interaction between stimulus order and filling, $F(1,14) = 7.8$, $p = .014$, as well as an interaction between stimulus pairs, stimulus order, and filling, $F(4,56) = 4.1$, $p = .006$. We shall consider this effect in greater detail below.

Discussion

The results of Experiment I support the hypothesis that isolated steady-state vowels will be perceived categorically when there is interference with auditory memory. Discrimination performance in the long filled condition was close to that predicted under strict categorical perception assumptions. It seems likely that the combined effects of decay and interference in this condition impaired the auditory trace of the first stimulus in a pair to a degree that made an auditory comparison with the second stimulus rather difficult. Consequently, the listeners probably relied on phonetic memory codes in the most difficult condition, whereas, in the easier conditions, phonetic memory was supplemented by varying amounts of auditory memory.

Our results show that auditory memory is vulnerable to both decay and interference. The fact that performance in the long unfilled condition was better than in either of the filled conditions suggests that the auditory memory for the first stimulus in an AX pair took longer than 2 sec to decay, given that no interfering sounds followed.

The parallelism of the discrimination functions in Figure 1 is somewhat surprising. One might have expected that between-category comparisons, which -- according to the dual-coding model -- need not rely much on auditory memory, would be less affected by interference than within-category comparisons, which presumably rely more on auditory memory. (Of course, the distinction between between- and within-category comparisons is not an absolute one, considering the high frequency of confusions -- see Table 2.) In other words, the twin peaks in the discrimination function might have been expected to be most pronounced in the long filled condition but flatter in the short unfilled condition, where performance approached the ceiling. It would be implausible to assume that phonetic memory was affected by our manipulations of the ISI. However, the results can be explained by the unequal spacing of the stimuli. It will be recalled that there was a larger acoustic distance between categories than within. Thus, the contribution of auditory memory to between-category comparisons was increased, and this may explain why the peaks in the discrimination function remained pronounced even in the short unfilled condition. According to this interpretation, the peaks in the long filled condition reflect the higher phonetic discriminability of between-

category comparisons, whereas the similar peaks in the short unfilled condition reflect the higher auditory discriminability of these same comparisons.

This confounding of auditory distance with phonetic distance was an unintended consequence of our attempt to choose stimuli that were maximally representative of their respective categories. Now it could be argued that our listeners did not use phonetic categories at all but made discriminations exclusively on the basis of auditory stimulus codes -- a hypothesis that would be congenial to several recent discussions of categorical perception (Carney et al., 1977; Macmillan, Kaplan and Creelman, 1977). The peaks in the discrimination functions, it might be argued, represent simply the superior auditory discriminability of the between-category comparisons (stimulus pairs 3-6 and 8-11); the remaining peaks in the long filled condition reflect residual auditory memory for these larger stimulus differences, and their agreement with the predictions (Figure 1) is purely coincidental. Most likely, this view could be rejected by showing that, if even more severe interference with auditory memory is introduced, performance does not deteriorate further but remains at the level of the present long filled condition, and thus in accordance with the predictions based on phonetic labeling. Alternatively, an experiment with equally spaced stimuli might be conducted, in order to unconfound auditory and phonetic distinctiveness. We chose the latter course in our second experiment, which was designed to provide considerably more detailed data on the relationship between labeling and discrimination.

EXPERIMENT II

Experiment II employed a 13-member vowel continuum (Table 1), in which the stimuli were separated by nearly equal logarithmic steps. There was only one interference condition, corresponding to the long filled condition of Experiment I, and a short unfilled condition. As in the previous experiment, an identification test was included in order to predict discrimination performance and thus to test whether perception was categorical in the long filled condition.

However, the present experiment included an important new feature. In addition to obtaining discrimination responses to the AX pairs in the short unfilled and long filled conditions, we also asked the subjects, in two separate conditions, to give phonetic labels to the stimuli in exactly the same AX pairs. This provided us with information about the subjects' choice of labels as a function of the surrounding stimulus context, and with a new, and probably more appropriate, set of predictions to be compared to actual discrimination performance. The reason we expected these new "in-context" predictions to be more appropriate than those derived from a single-item identification test is the well-known fact that vowel identification is affected by the surrounding context, usually in the form of contrast (Fry et al., 1962; Eimas, 1963; Lindner, 1966; Thompson and Hollien, 1970; Ainsworth, 1974; Kanamori, Kasuya, Arai and Kido, 1971). By taking such contrast effects into account, we expected to obtain a more accurate estimate of the probabilities of the various labels that the subjects may have covertly applied in the discrimination task, and thus a more accurate estimate of the degree to which

discrimination responses might have been based on such labels. Apart from this comparison, we were interested in the contrast effects themselves as an object of study: how large they would be; whether they would occur in both directions in an AX pair (proactive and retroactive contrast); and whether they would be affected by the interfering sounds in the long filled condition.

We should emphasize at this point that we were interested in separating two questions that often have been treated in the past as the single issue of categorical perception: whether perception is absolute, and whether discrimination is based on phonetic labels. The original definition of categorical perception, as put forth by Liberman et al. (1957) and Studdert-Kennedy et al. (1970), implies absolute, that is, context-independent perception. If the labeling of a stimulus depends on the preceding or following stimuli, as frequently seems to be the case with vowels, perception is by definition noncategorical. However, the listeners may nevertheless use these context-dependent categories in a discrimination task. Thus, it may be necessary to distinguish two kinds of noncategorical perception: one in which discrimination is based on a few discrete categories, and one in which perception is truly continuous -- that is, not mediated by categorization.

To illustrate these distinctions -- categorical, category-based noncategorical, and continuous perception -- consider the two methods of prediction employed in Experiment II. One set of predictions is derived from the labeling responses in a single-item identification test. These predictions provide an unambiguous test of categorical perception. In order for the discrimination results (in the long filled condition) to match these predictions, it is necessary not only that the subjects base their discrimination responses on the same labels as in the identification task, but also that the probabilities of applying these labels be the same in the two situations. Since the stimulus contexts are quite different in the single-item test and in the AX test, this equality of labeling probabilities necessarily requires that the process of categorization be context-independent. Experiment I suggested that this might be true in the long filled condition. Experiment II provided a much more stringent test of this hypothesis, since it included a direct comparison of labeling behavior in different contexts. It is essential to realize that the equality of the labeling probabilities in the single-item and AX tests is a necessary condition for the perception of the vowels to be called categorical (see Studdert-Kennedy et al., 1970).

Consider now the case in which the labeling probabilities are not the same in these two tests. This situation will hold almost certainly in the unfilled short condition, where contrast effects are expected; it may also turn out to apply to the long filled condition. It is here that the second set of predictions comes in -- predictions derived from the labeling probabilities in AX context. Context dependence implies that this set of predictions will be different from that derived from the single-item identification test. Moreover, there are separate sets of AX predictions for the long filled and short unfilled conditions, and they may also be different from each other, to the extent that the labeling probabilities vary between these two conditions. However, if these in-context predictions match the discrimination results obtained in corresponding conditions, it may be argued that discrimination is mediated by (context-dependent) category labels, despite the fact that percep-

tion is not categorical. Only if discrimination performance matches neither set of predictions, could we be convinced that perception is truly continuous.

Thus, we distinguish three different possibilities: (1) perception is truly categorical, and both sets of predictions (those derived from single-item identification and those derived from AX labeling) coincide with the discrimination results; (2) only the in-context predictions (derived from AX labeling) are matched by the discrimination results, suggesting that discrimination may be mediated by discrete categories; (3) the discrimination results match neither set of predictions, and perception is truly continuous.[1]

Method

Subjects. Sixteen new volunteers participated. They were Yale undergraduates who received course credit for their participation.

Stimuli. The vowel continuum included all 13 stimuli listed in Table 1, a subset of which had been used in Experiment I. Three experimental tapes were prepared. The single-item identification tape contained a random sequence of 130 stimuli (10 repetitions of each of the 13 stimuli) with ISIs of 3 sec. Each of the other two tapes contained five different random sequences of 35 vowel pairs consisting of each stimulus paired with itself (13 pairs) and with every other stimulus two steps removed on the continuum, in both stimulus orders (22 pairs). One of the discrimination tapes had a short (300 msec) unfilled interval between the stimuli in each AX pair; the other tape had a long (1,920 msec) interval, filled with five repetitions of the /y/ sound, exactly as in the corresponding condition of Experiment I. The interpair interval was 4 sec, and blocks of 35 pairs were separated by an extra 4 sec.

Procedure. All subjects first took the single-item identification test. As in Experiment I, they circled "beet," "bit," or "bet" on an answer sheet. This task was followed by the two AX tapes presented twice with different instructions. Under discrimination instructions, the subjects circled "s" (same) or "d" (different) on the answer sheet, as in Experiment I. Under labeling instructions, the subjects circled "beet," "bit," or "bet" for each of the two vowels in a pair. (On the answer sheet, the three response

---

[1]To the best of our knowledge, we are the first actually to compute in-context predictions in a categorical-perception task, although Lane (1965) suggested the idea long ago. Several of the earlier studies on vowel perception (Eimas, 1963; Fry et al., 1962; Fujisaki and Kawashima, 1969, 1970) obtained labeling responses to the precise stimulus sequences used in discrimination, but none of these studies made the predictions conditional on context. Instead, all labeling responses were lumped together, thus averaging out all context effects. Most likely, this accounts for the large discrepancies between predicted and obtained discrimination performance, particularly in the often-cited study by Fry et al. (see Lane, 1965). It is not clear why these authors used the discrimination sequences to collect labeling responses in the first place.

alternatives appeared twice side by side.) The subjects were instructed to listen to both vowels before responding. The sequence of the discrimination and labeling conditions was counterbalanced across subjects, and so was the sequence of the short unfilled and long filled conditions within each instruction condition.

Results

Simple Identification. The results of the single-item identification test are summarized in Figure 2. The percentages of responses in the three categories, /i/, /I/, and /ɛ/, are shown as a function of stimulus location along the continuum. As in Experiment I, and in agreement with Pisoni's (1971) results, the stimuli were less consistently assigned to the middle category, /I/, than to the other two categories.

As in Experiment I, we used these identification results to predict discrimination performance. The resulting predictions, averaged over subjects, are represented by the dashed function at the bottom of Figure 3. The function has two peaks, reflecting the prediction of higher discrimination performance in the category boundary regions. If vowels are categorically perceived in the absence of auditory memory, the discrimination results in the long filled condition should coincide with these predictions.

AX Discrimination. The results of the discrimination task are displayed in Figure 3 in terms of percent correct responses, derived and plotted in the same manner as in Experiment I (solid functions). Performance in the short unfilled condition was much better than in the long filled condition, as expected, $F(1,14) = 78.8$, $p < .0001$. Discrimination performance also varied significantly with location on the stimulus continuum, $F(10,140) = 7.8$, $p < .0001$; there was a pronounced peak in the region of stimulus 4. As in Experiment I, the short unfilled and long filled discrimination functions were strikingly parallel, as confirmed by the absence of a significant interaction between the factors of ISI and location on the continuum, $F(10,140) = 1.4$, $p = .186$). This suggests that the parallelism of the discrimination functions in Experiment I was not an artifact of unequal stimulus spacing along the physical continuum.

In Experiment I, discrimination performance in the long filled condition resembled the predictions derived from single-item identification performance. However, the data of the present experiment do not support this earlier observation. Although predicted and obtained performance were close in the middle range of the continuum, the obtained scores were clearly better than predicted at the ends of the continuum, particularly at the right (/ɛ/) end. For all but three of the subjects obtained performance exceeded predicted performance in the long filled condition, and Chi-square tests on individual subjects revealed that 8 of the 16 subjects performed significantly better than predicted ($p < .05$). Even more importantly, the shape of the obtained discrimination function did not conform to the predictions. Specifically, the predicted peak in the /I/-/ɛ/ boundary region was absent, and the predicted peak in the /i/-/I/ boundary region was displaced to the left. These discrepancies could be explained by assuming that the labeling probabilities of the stimuli changed in the context of the AX pairs. Therefore, the

predictions derived from the single-item identification test were not appropriate, and, hence, perception was not truly categorical in the long filled condition. In Experiment I, the wider and unequal stimulus spacing forced similar zigzag shapes on both predicted and obtained functions, and thus prevented us from detecting any serious mismatch. We proceed now to a discussion of the AX labeling results that were expected to provide more accurate predictions of discrimination performance, since they were obtained in identical presentation contexts.

AX Labeling

The predictions derived from the AX labeling responses are shown in Figure 3 as the two dotted functions. We computed predicted percent correct discrimination scores, considering each pair of AX labeling responses placed in the same phonetic category as equivalent to a "same" response and each pair of responses placed in different phonetic categories as equivalent to a "different" response. If it had been true that each vowel was identified independently of its context, the predictions from the AX labeling task at both ISIs should have equalled the predictions from the single-item identification test. This was clearly not the case, not even in the long filled condition, thus providing indirect evidence for context effects in labeling.

The discrimination scores derived from the AX labeling task (the in-context predictions) were much closer to the results of the discrimination task than to the predictions from the single-item identification test. Like discrimination, AX labeling performance showed a strong effect of interference $F(1,12) = 52.6$, $p < .0001$, and of location on the continuum $F(10,120) = 13.3$, $p < .0001$. There was a small interaction between these two factors, $F(10,120) = 2.9$, $p = .003$; however, the functions for the short unfilled and long filled conditions were essentially parallel. They were also similar in shape to the functions obtained under discrimination instructions, showing only a single peak at stimulus 4.

The in-context predictions represent the discrimination performance to be expected when only the prescribed phonetic category labels are used. However, the scores actually obtained in the discrimination task, while similar in profile, significantly exceeded these expectations, $F(1,12) = 11.7$, $p = .005$. Figure 3 shows that this difference derived from the two ends of the vowel continuum, particularly the right (/ɛ/) end, while scores in the middle region were similar. This was reflected in a significant interaction of task and stimulus location, $F(10,120) = 3.7$, $p = .002$. Especially interesting is the fact that the advantage of discrimination over labeling responses was as large in the long filled condition as in the short unfilled condition, as confirmed by a nonsignificant interaction of task and ISI, $F(1,12) < 1$.

Hits and False Alarms. In order to understand the results in more detail, we examined separately the responses to pairs of identical and pairs of nonidentical stimuli. Figure 4 plots the percentages of "different" responses to each stimulus pair, separately for the discrimination and labeling tasks. (In the labeling task, a "different" response represents a pair of responses placed in two different phonetic categories.) The two pairs

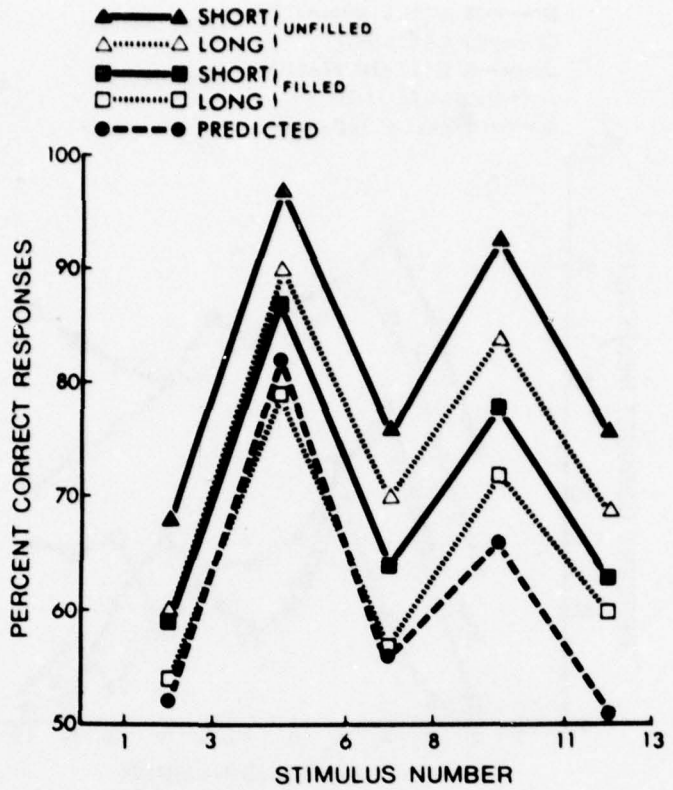


Figure 1: AX discrimination scores in the four conditions of Experiment I, together with scores predicted from identification responses.

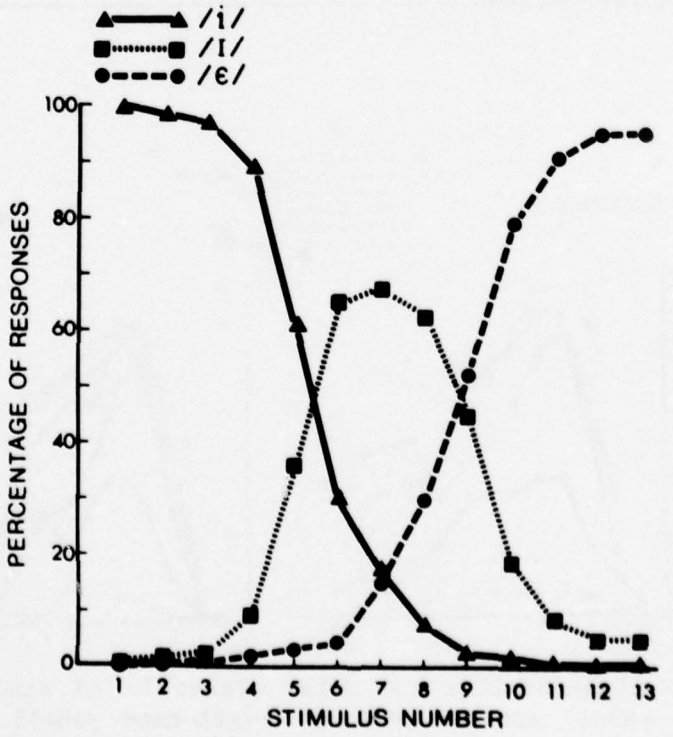


Figure 2: Labeling functions for stimuli presented in the single-item identification test of Experiment II.

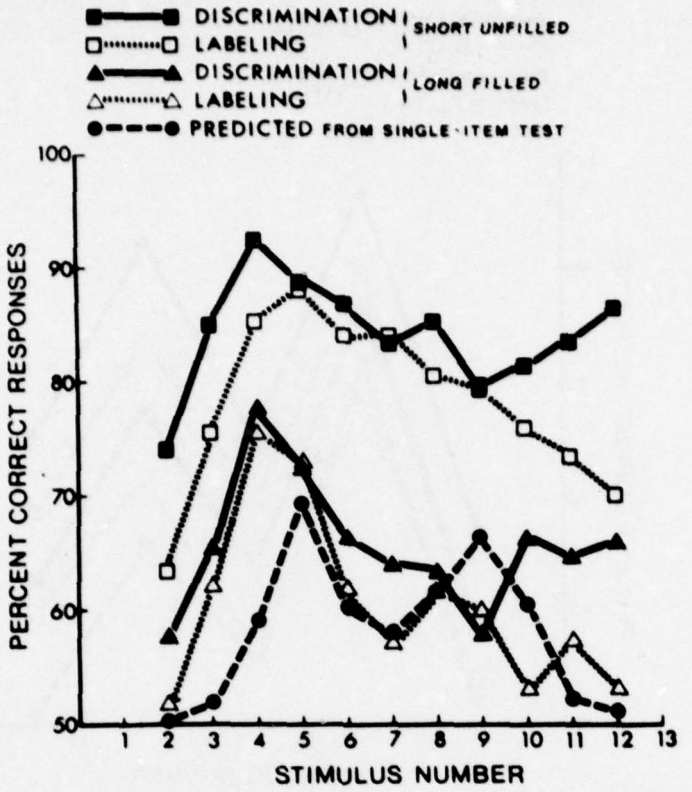


Figure 3: Two-step AX discrimination scores in the short unfilled and long filled conditions under discrimination and labeling instructions. The labeling results represent the "in-context" predictions of discrimination performance. Also shown are the predictions derived from the responses in the single-item identification test.

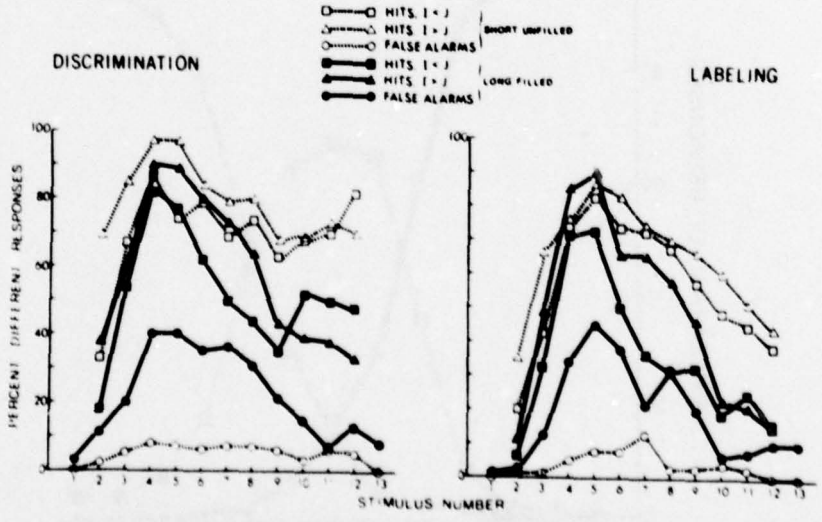


Figure 4: Percentages of hits and false alarms in AX discrimination (left-hand panel) and AX labeling (right-hand panel). Hit percentages are shown separately for the two different orders, I<J and I>J, of each pair (I,J), where I is the first and J the second stimulus.

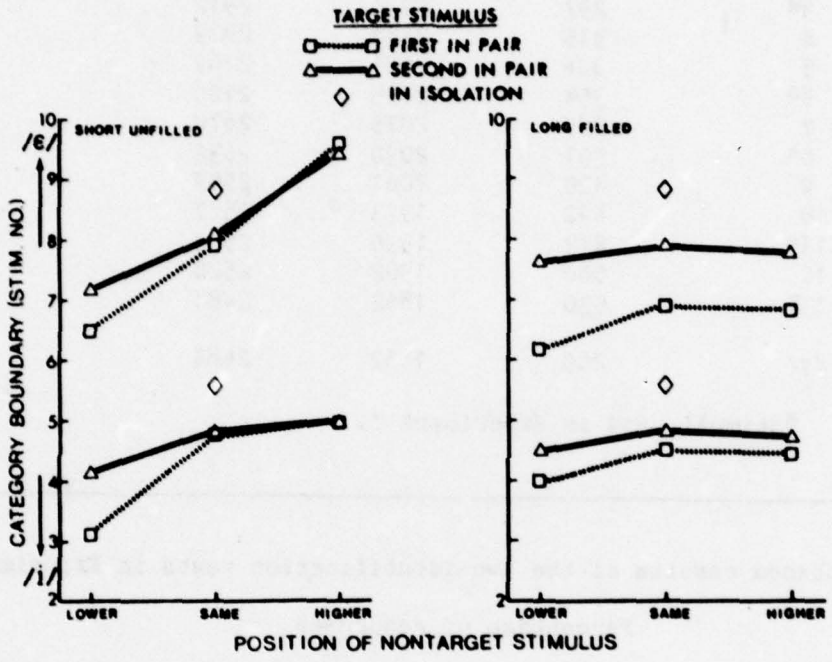


Figure 5: Category boundaries between /i/ and /I/ (bottom) and between /I/ and /ɛ/ (top) as a function of the relative position of the nontarget stimulus, shown separately for the first and the second stimulus in a pair as the target (retroactive vs. proactive contrast). The boundaries for the stimuli in the single-item identification test are also shown ("in isolation").

TABLE 1: Formant frequencies of the vowel stimuli (in Hz).

Stimulus	F_1	F_2	F_3
1[a]	269	2296	3019
2	285	2263	2955
3[a]	297	2230	2912
4	315	2183	2829
5	336	2151	2769
6[a]	354	2105	2709
7	375	2075	2670
8[a]	397	2030	2632
9	420	2001	2567
10	442	1973	2557
11[a]	472	1930	2539
12	500	1902	2520
13[a]	530	1862	2484
/y/	269	1862	2484

[a]Stimuli used in Experiment I.

TABLE 2: Combined results of the two identification tests in Experiment I.

	Percentage of responses		
Stimulus	/i/	/I/	/ɛ/
1	99	1	0
3	90	8	2
6	9	80	11
8	1	60	39
11	0	11	89
13	0	4	96

TABLE 3: Stimulus order effect in Experiment I.

Stimuli	Percentage of "different" responses	
	i < j	i > j
1, 3	21	43
3, 6	93	97
6, 8	34	63
8, 11	63	79
11, 13	41	34

Note: i = first stimulus; j = second stimulus.

Table 4: AX labeling task: Percentages of /i/ and /ɛ/ responses as a function of position of target stimulus in pair (first or second), relative location of nontarget stimulus (lower or higher), and ISI.

	Position of target stimulus			
	First		Second	
Interval	/i/	/ɛ/	/i/	/ɛ/
Short unfilled				
Lower	14	55	20	49
Higher	28	27	28	30
Long filled				
Lower	18	60	22	45
Higher	21	53	25	42

of functions at the top of each panel represent "hits," that is, "different" responses to pairs of different stimuli, at the two ISIs. The difference between the two functions at each interstimulus interval may be ignored for the moment. The two functions at the bottom of each panel are "false alarms," that is, "different" responses to pairs of identical stimuli, at the two ISIs.

Although the percent-correct discrimination functions for the two ISIs had been quite parallel (Figure 3), this was not true for hits and false alarms considered separately. In both, the effect of interval interacted with stimulus location, in a complementary fashion: whereas hits showed strong effects of interference at the ends of the continuum only, false alarms showed large effects in the middle of the continuum only. This was true regardless of the task performed. The interaction of the interference effect with location on the continuum was significant for both hits, $F(10,120) = 6.6$, $p < .0001$, and false alarms $F(12,180) = 6.7$, $p < .0001$, and there were no interactions of these factors with task. Thus, at the long filled interval, discrimination errors in the middle of the continuum were largely due to false alarms, while errors at the ends of the continuum were largely misses. Another way of expressing this result is that, in the presence of interference, the subjects were more likely to respond "different" in the middle of the continuum than at the ends. This may have been a consequence of the general uncertainty about the middle category, /I/ (see Figure 2).

An analysis of variance showed that there were significantly more hits in the discrimination task than in the labeling task, $F(1,12) = 15.5$, $p = .002$; this difference derived primarily from the right end of the stimulus continuum, leading to a significant interaction of task with stimulus location for hits, $F(10,120) = 4.7$, $p < .0001$. On the other hand, the false-alarm rates did not differ between the two tasks (see Figure 4). Thus, the higher scores in the discrimination task were due to a higher hit rate, not a lower false-alarm rate.

Stimulus Order Effect. The strong stimulus order effect obtained in the discrimination task of Experiment I was replicated in the present study. Figure 4 displays two hit functions for each ISI in each task. These two functions are distinguished only by the order of the stimuli in a pair. The functions connecting squares represent the order I<J, where the first stimulus in a pair (I) had a lower position on the continuum than the second stimulus (J); the functions connecting triangles represent the reverse order, I>J. It can be seen that the majority of stimulus pairs received more "different" responses in the order I>J than in the order I<J, but this effect disappeared or was even reversed at the right end of the continuum. This pattern of results was reflected in a significant interaction of stimulus order and location, $F(10,120) = 2.9$, $p = .003$, together with a significant main effect of stimulus order, $F(1,12) = 8.8$, $p = .012$. The stimulus order effect was present in both tasks and, most interestingly, at both ISIs. In the middle of the continuum, the effect was actually increased by interference, which contributed to a significant three-way interaction involving stimulus order, location, and interference, $F(10,120) = 2.8$, $p = .004$. This contrasts with the results of Experiment I, where a small decrease in the stimulus order effect was observed as a function of interference. Taken together, however, the two findings justify the conclusion that the stimulus order effect was

little affected by interference.

Contrast Effects. The results of the AX labeling task offered an opportunity to investigate the degree to which the relationship between the stimuli in a pair influenced identification. Two effects were of special interest: whether the (expected) contrast effect would be stronger in one direction than in the other (proactive vs. retroactive contrast), and whether its magnitude would change as a function of interference with auditory memory.

In order to answer these questions, we first tabulated the labeling response frequencies in the three phonetic categories separately for stimuli occurring first and stimuli occurring second in pairs of different stimuli, and then examined these frequencies for one (target) stimulus contingent on the nature of the other (nontarget) stimulus in the pair. The nontarget stimulus could be either lower on the continuum (-2 steps), identical to the target, or higher on the continuum (+2 steps). For each of these three cases, perceptual boundaries between adjacent vowel categories were determined from the average data using Finney's (1971) probit algorithm. This procedure provides an estimate of the 50 percent crossover point of the labeling functions for adjacent categories, that is, of the category boundary. The two boundaries--that between /i/ and /I/ and that between /I/ and /ɛ/--were expressed in terms of their location on the stimulus continuum. Figure 5 shows these boundaries as a function of the relative position of the nontarget stimulus, separately for the first and second stimulus as target, and with separate panels for the short unfilled and long filled intervals. In addition, the boundaries obtained in the single-item identification test are shown (in isolation, see Figure 2). The functions on top represent /I/-/ɛ/ boundaries, while those at the bottom represent /i/-/I/ boundaries. (Note that stimulus location, plotted on the abscissa in previous figures, is plotted on the ordinate in Figure 5.)

Since there were too few observations to compute reliable boundaries for individual subjects, an analysis of variance was conducted on response percentages pooled over all pairs in a given condition, with the following factors: vowel category (/i/ vs. /ɛ/; /I/ responses were omitted), target stimulus (first vs. second), relative location of nontarget stimulus (higher vs. lower; identical pairs were not included in this analysis), and interference. Pairs including target stimuli 1, 2, 12, and 13 were omitted since these stimuli could not be paired with both higher and lower stimuli on the continuum. These response percentages, which formed the basis of the statistical tests, are shown in Table 4, averaged over subjects.

A contrast effect implies a shift in the category boundaries for target stimuli toward the category represented by the nontarget stimulus. In other words, if there is a contrast effect, the category boundaries for target stimuli will be shifted toward the lower end of the continuum when the nontarget stimulus has a lower position, and they will be shifted toward the higher end when the nontarget stimulus has a higher position on the continuum. This implies a positive slope for the "boundary functions" (the connected points) in Figure 5. Obviously, there were pronounced contrast effects in the short unfilled condition (left panel), but only negligible effects in the long filled condition (right panel). The overall contrast effect was significant,

$F(1,15) = 23.0$, $p = .0003$, as well as its interaction with the interference factor, $F(1,15) = 19.8$, $p = .0005$. In a separate analysis of the long filled condition, the contrast effect still reached significance, $F(1,15) = 5.1$, $p = .04$, although it was obviously very small. A slope difference between the solid and dashed functions in Figure 5 would reflect a difference between proactive contrast (second stimulus as target) and retroactive contrast (first stimulus as target). It can be seen that, surprisingly, the retroactive effect was slightly stronger than the proactive effect at the unfilled short interval, although this difference turned out not to be significant.

A level difference between the solid and dashed functions in Figure 5 implies a boundary shift as a function of stimulus position in an AX pair. Such a difference can be observed in the long filled condition, $F(1,15) = 10.6$, $p = .006$; it suggests that the perception of the vowels may have been influenced by the five interpolated /y/ sounds, although the control identification test in Experiment I had shown no effect of a single /y/ precursor on labeling. Five repetitions of /y/ may have been sufficient to produce selective adaptation (Morse, Kass, and Turkienicz, 1976) which, of course, is a kind of contrast effect. The fact that the category boundaries shifted toward the lower (/i/) end of the continuum is in agreement with this interpretation and with the intuitive observation that /y/ is perceptually more similar to /i/ (with which it shared the first formant) than to /ɛ/ (with which it shared the weaker, higher formants). Curiously, however, the data in Figure 5 (right panel) indicate that it was the vowel preceding the five /y/ stimuli (the first vowel in an AX pair) whose boundary shifted the most, not -- as one might expect in an adaptation situation -- the following vowel. Thus, these boundary shifts remain somewhat puzzling.

Apart from any differential effects, the boundaries in the AX labeling task were generally shifted towards the lower (/i/) end of the continuum, relative to the boundaries for the same stimuli in isolation. This is particularly evident when the latter are compared with the boundaries for identical AX pairs, in which the same stimulus was repeated once. It is not known what caused this shift in perception, but it obviously contributed to the discrepancy, shown in Figure 3, between the predictions derived from the AX labeling task and those derived from the single-item identification test. This finding demonstrates that stimulus context may affect the labeling probabilities even in the absence of contrast effects (that is, in pairs of identical stimuli).

Discussion

Are Vowels Perceived Categorically? The principal question of our research was whether isolated steady-state vowels would be perceived categorically when there is interference with auditory memory. Experiment I suggested an affirmative answer. However, the much more fine-grained analysis in Experiment II indicated that the answer depends on the exact form in which the question is asked. Recall that in the introduction to Experiment II, we distinguished among three modes of perception--categorical, category-based noncategorical and continuous. We have obtained a fairly close fit between the in-context predictions and obtained discrimination performance. (A small surplus of discrimination performance over the predictions is a common finding

even with stop consonants, and although it requires an explanation, it is not considered a major argument against categorical perception.) Such a reasonable fit between predicted and obtained discrimination functions -- without any qualifications about the nature of the identification test from which the predictions are derived -- has often been considered the sole criterion of categorical perception (see especially Macmillan et al., 1977). However, we have also found strong evidence for auditory contrast effects in vowel labeling, thereby indicating relative rather than absolute perception. These two results suggest that it is the second mode that is employed for vowels--noncategorical but category-based.

Thus, the answer to our original question depends entirely on how we choose to define categorical perception. By the predictability criterion (using the appropriate in-context labeling data) we succeeded in making vowel perception categorical when there was interference with auditory memory. In fact, even when auditory memory was intact, perception was categorical by this criterion. However, as we have pointed out earlier, categorical perception--as defined by Liberman et al. (1957)--is synonymous with absolute, context-independent perception. ("Absolute" is, incidentally, the primary dictionary definition of the word "categorical.") Therefore, any evidence indicating that labeling behavior depends on stimulus context argues against categorical perception. We obtained such evidence: in the case when the two test stimuli were close together, there were reciprocal contrast effects. When the two stimuli were separated by time and interference, reciprocal contrast effects in labeling were minimal. However, the labeling probabilities in this latter case nevertheless deviated considerably from those in the single-item identification test. This was probably due in part to contrast with the interpolated /y/ stimuli. The mere existence of phonetic category boundary shifts as a consequence of changes in stimulus arrangement indicates that the stimuli were not perceived absolutely. Also, the labeling probabilities depended on the absolute position of a stimulus in an AX pair -- the stimulus order effect. Thus, it appears that some stimuli, because of their particular acoustic structure, are perceived relative to the surrounding context, and cannot be made to be perceived absolutely by manipulating that context. Rather, any new context -- such as the /y/ vowels introduced to interfere with auditory memory -- will simply constitute a new frame of reference for the relative perception of the target stimuli.

Vowels, consonants, and the operational definition of categorical perception. Although we have made no direct comparisons between performance on vowels and on stop consonants, our results suggest some similarities and differences. By the predictability criterion, based on in-context identification data, vowels and stops are not likely to be very different in view of the high degree of predictability we observed here for vowels. Alhough we did obtain a significant discrepancy between predictions and discrimination performance for vowels, it is well known, as we observed earlier, that there is also, typically, a small discrepancy for stops. An interesting possibility, which we are presently testing, is that the discrepancy for stop consonants could be reduced considerably by basing the predictions on in-context identification rather than on single-item identification, as has been done previously. It is known that even stop consonants show small context effects (Eimas, 1963), and the in-context prediction procedure would take such effects into

account. However, even though the fit between such predicted and obtained discrimination may turn out to be somewhat closer for consonants than for vowels, the fact remains that predictability is high for both kinds of stimuli. The fundamental difference between stop consonants and vowels seems to lie in their degree of susceptibility to context effects in identification. These effects seem to be larger for vowels than for consonants.

The traditional definition of categorical perception has included two aspects, absoluteness in phonetic labeling and predictability of discrimination performance from labeling performance. However, only the predictability requirement was ever directly operationalized. The other requirement--that labeling be context-independent--was generally satisfied by the wide stimulus spacing in the single-item identification test. Meeting the traditional operational definition (predictability of discrimination from single-item labeling data) does indeed indicate categorical perception, but failure to meet it is ambiguous. Lack of fit between identification and discrimination could be caused by a failure in either or both aspects of the definition. We prefer two separate tests, one for absoluteness and one for the predictability. Both tests make use of in-context labeling performance. Absoluteness is indexed by the effects on labeling of stimulus context, and the predictability measure is strictly analogous to the traditional one, except that in-context predictions are applied. Because interesting differences between vowels and stop consonants seem more likely to be found on the absoluteness test, that may turn out to be the more informative part of the new operational definition.[2]

The Roles of Auditory and Phonetic Processes in Vowel Discrimination. Having commented on the degree of categorical perception in vowels, we now wish to discuss the processing mechanisms that our subjects may have brought to bear on the AX discrimination task. According to the conventional logic, meeting the predictability requirement of categorical perception directly implies a process account of discrimination performance: the subject bases his response entirely on phonetic labels. However, even though our results demonstrated such predictability, there are several permissible process explanations.

The results to be explained. There are five findings that should be considered in any comprehensive process account:

---

[2]In our discussion, we have more or less ignored two other criteria for categorical perception commonly cited in the literature: the relative steepness of the labeling functions, and the presence of peaks and troughs in the discrimination function. We find these criteria less important because they are more difficult to quantify than the fit between predicted and obtained discrimination, and because they are more or less directly related to context effects in categorization. The relatively shallow slopes of the labeling functions (see Figure 2) and the irregularities in the discrimination functions in Experiment II generally support our conclusion that the perception of the vowels was not absolute.

(1) The AX discrimination performance was quite well predicted by in-context labeling performance, although there was a statistically significant difference due primarily to a discrepancy at the /ɛ/ end of the vowel continuum we used.

(2) As is usual with speech stimuli, we obtained discrimination peaks approximately at the category boundaries, although there was only a single peak in Experiment II.

(3) Discrimination was poorer after a long filled interval between AX stimuli than after a short unfilled one.

(4) There were large reciprocal contrast effects in the AX labeling task at the short unfilled interval; these were greatly reduced at the long filled interval.

(5) There were clear stimulus order effects in discrimination that were not consistently a function of interference or of delay.

The Fujisaki-Kawashima-Pisoni dual-coding model. Fujisaki and Kawashima (1969, 1970) and Pisoni (1971, 1973, 1975) have offered a process model for discrimination performance of speech stimuli. The main assumption of this model is that there are two codes that may be used to make comparisons of stimuli--phonetic and auditory memory. Whenever two stimuli cannot be distinguished by their phonetic codes, the listener is assumed to consult his auditory memory code. It follows that the differences between predicted and obtained discrimination performance, presumably even the small discrepancies obtained here with the in-context predictions, are due to the contribution of auditory memory. This model falsely predicts that in our short unfilled condition, where there should have been abundant auditory information about the first stimulus at the time of arrival of the second stimulus, the predicted-obtained discrepancy should have been considerably larger than in the long filled condition, where little auditory information should have survived. Instead, we found equally small discrepancies in the two conditions. Thus, our results strongly contradict the predictions of the dual coding model.

An all-auditory model for discrimination. One model that can deal successfully with our results is based on the assumption of a single auditory memory code for comparing the two stimuli. This model deals more or less successfully with each of the five results listed above. Deterioration of discrimination performance following a long filled interval is an obvious consequence of this model because auditory memory is assumed to deteriorate with time. The disappearance of contrast effects upon labeling with a long filled ISI is also consistent for the same reason. These contrast effects may have a sensory basis similar to that presumed to underlie brightness contrast in vision. Alternatively, contrast could be caused by the conscious strategy of giving different phonetic labels to two sounds whenever they sound different. In other words, phonetic categorization in the labeling task may be strongly influenced by the result of implicit auditory discrimination judgments.

The all-auditory model may account for discrimination peaks only on the basis of some acoustic discontinuity corresponding to category boundaries (Pastore et al., 1977). This model assumes no phonetic processing during AX discrimination, an assumption that makes it difficult to accommodate the close fit between discrimination performance and predictions based on phonetic labeling. It may be, however, that the relatively small range of the present stimulus continuum enabled the subjects to achieve relatively high resolution in labeling with only a small number of phonetic categories (see Pynn, Braida and Durlach, 1972; Ades, 1977). On the other hand, discrepancies between obtained discrimination and predictions are no problem for this model and, indeed, we obtained such a discrepancy, especially at the /ɛ/ end of our vowel continuum.

Two of our results provide some difficulty for the all-auditory model. First, performance in the long filled condition was well above chance. This means that there must have been some substantial auditory memory persisting over the long filled interval. The implication is that if we had used a more effective delay interval (or interference stimulus) the subjects either would nave been left performing at chance or would have had to adopt a different processing strategy. More serious is the difficulty in accounting for the stimulus order effect on an auditory basis. This order effect was not consistently affected by time delay or interference, which do affect performance assumed to reflect auditory memory.

An all-phonetic model for discrimination. Another model that can deal successfully with our results is based on the assumption of a single phonetic code as the basis for comparisons. This model does not deny a role for auditory memory in discrimination: the discrimination responses are based entirely on phonetic distinctions, but those phonetic labels have themselves been subject to auditory influences. That is, phonetic coding occurs first on the basis of auditory information but it is only these codes that are then used for discrimination.

The close fit between discrimination and predictions based on phonetic labeling is the natural outcome of this model. According to this model there is no difference in the subject's information processing in the two tasks, only that the labels are covert in one case and overt in the other. Any surplus discrimination over that predicted by labeling must be explained by the presence of additional covert phonetic categories used in discrimination but ineligible for the labeling task (see Chistovich and Kozhevnikov, 1970, for a similar argument); these additional categories should be equally available whatever the interference condition. Figure 3 shows that the obtained surplus occurred precisely in our stimulus continuum where there is reason to believe that an extra covert category did exist.[3]

---

[3]Informal evidence suggests that there may have been an additional phonetic category, /e/, in the region of the /I/-/ɛ/ boundary. Since /e/ is not an English phoneme (the diphthong /e^{I}/ occurs instead), it was not included among the response alternatives. Some subjects may have made covert use of this additional category in the discrimination task and thus may have widened the gap between predicted and obtained discrimination.

The occurrence of discrimination peaks located at category boundaries is another direct consequence of the all-phonetic model. The fact that only one peak was found in the discrimination task of Experiment II is no problem for the all-phonetic model since only one peak was found in the AX labeling task of Experiment II, and subjects are necessarily basing their responses on phonetic codes in that task.

In order for the all-phonetic model to account for the poor discrimination obtained at the long filled interval, hits and false alarms should be considered separately: the probability of hits (correctly saying "different" to physically nonidentical stimuli) is increased at the short unfilled interval by reciprocal contrast between the two stimuli, which is greatly reduced at the long filled interval. False alarms (incorrectly saying "different" to physically identical tokens), according to the all-phonetic model, result from inconsistency in the labeling of a given stimulus token. It is natural to expect that momentary fluctuations in the subject's state would produce increasing inconsistency as the two stimuli are more and more separated.

Another explanation of the performance level differences at the two intervals relies on the assumption that the subject waits to apply phonetic labels until both tokens have been received. Thus the first stimulus in the long filled condition will be in a state of degraded representation in auditory memory by the time the subject categorizes it. Therefore the phonetic label for this stimulus will have been based on poor information and both misses and false alarms will ensue. An internal test of this hypothesis is possible: in the long filled condition, the in-context labeling results should show steeper labeling functions for the second stimulus than for the first; there should be no such differences between the two stimuli in the short unfilled condition. The data partially bear out this hypothesis in that (a) there were essentially no differences in labeling for the two stimuli in the short unfilled condition, and (b) there were such differences in the long filled condition; however, (c) the latter differences existed only for the /I/ to /ɛ/ range of our continuum.

At present, the phonetic model provides no explanation of the stimulus order effect. However, the fact that the effect was equally large in both interference conditions suggests that a phonetic explanation may be appropriate.[4]

The all-phonetic model is designed to describe mechanisms resulting in the subject's "same" or "different" responses in discrimination; however, we may also expect from this hypothesis a statement on the contrast effects that were observed in in-context labeling. The form of this explanation is similar to one of the two all-auditory explanations of contrast effects: it is assumed that when two stimuli reside together in some stage of auditory processing, there are reciprocal interactions between the representations similar to those found in visual brightness contrast or other laterally

---

[4]See Appendix I.

inhibiting systems. At the short unfilled interval these processes are maximized in comparison to the long filled interval.

It might be objected that if there is a preliminary auditory stage in which contrastive effects occur, then why is not the AX discrimination response tapped directly off this contrastive process (rather than off the phonetic labels that, in turn, were influenced by the auditory contrast). Such direct use of the preliminary auditory stage would require consciousness of it, which, however, we consider to be unlikely. Conscious attention cannot be directed simultaneously to two different levels of analysis of the same message. When one level consists of a linguistically significant channel, it seems to be preferred. The Stroop color-word interference effect shows this rule operating in the visual realm; subjects cannot disregard the linguistically informative verbal channel while naming the colors of printed words. It is as if the linguistic level of analysis always dominates the nonlinguistic level; this is of obvious adaptive value for human communication. In our type of experiment, Bailey, Summerfield and Dorman (1977) have shown how difficult it is to leave the phonetic mode once subjects have begun to place phonetic interpretations on sounds at first perceived as nonspeech. The all-phonetic model for discrimination is in harmony with these considerations; phonetic mediation for purposes of discrimination is a natural and automatic consequence of inherent priority for the linguistic level of analysis when one exists in stimuli. Of course, the whole process of discrimination is fundamentally guided by auditory information processing. The information, after all, enters the system through the ears. The question at issue is whether or not there is a stage of phonetic mediation from which discrimination responses are drawn. By arguing in favor of this notion, we do not exclude the possibility that the subjects' attention may be directed to the auditory level by extended practice or special discrimination paradigms; this has been shown to be possible even with stop consonants (Ganong, 1977; Samuel, 1977). However, we find it plausible to assume that the relatively inexperienced listeners in our experiments followed a natural tendency to remain in the phonetic mode of processing.

A mixed model for discrimination. It is possible to combine the assumptions of the all-auditory and all-phonetic models for discrimination into a mixed model. It postulates that auditory processing dominates at the short unfilled interval and phonetic processing dominates at the long filled interval. The problems raised earlier for the all-auditory model apply with equal force to this hybrid model. Additionally, there is the unique problem associated with the hybrid model that the near identical goodness of fit between predicted and obtained discrimination levels is observed in the two conditions, which must be assumed to be purely coincidental. Similarly, the stimulus order effects were similar in the two conditions but would be explained by the hybrid model as resulting from two different mechanisms in the two conditions. Likewise, the obtained discrimination peak has to be explained in one way for the short unfilled condition and in another way for the long filled condition. Thus, despite the greater flexibility of the mixed model, it seems clearly unparsimonious compared to the others.

Conclusions

We remain in some doubt as to the detailed processing model that supports the AX discrimination of isolated vowels. However, this uncertainty should not detract from the positive conclusions permitted by our experiments:

Phonetic labeling is an excellent predictor of AX discrimination performance provided that the labels are obtained in the same context that is used in discrimination testing. This was true even under conditions presumed to be rich in auditory memory. Thus, the original version of the dual coding model for speech discrimination (the Fujisaki-Kawashima-Pisoni model) needs to be revised.

Reciprocal (proactive and retroactive) contrast effects are a major influence on phonetic labeling of vowels. It is on the basis of this evidence that we conclude that vowels are not perceived categorically. Nevertheless, it appears that vowel discrimination may be mediated by phonetic labels, possibly even to the same extent as is discrimination of stop consonants.

APPENDIX I

An interesting hypothesis was proposed by Smith* who apparently was the first to discover the stimulus order effect with vowels. She refers to the time-order error often found in studies of duration discrimination (see, for example, Jamieson and Petrusic, 1975) and links this finding with the fact that /I/ tends to be shorter than /i/ and /ɛ/ in natural speech (Peterson and Lehiste, 1960). Thus, stimulus duration provides an additional cue for distinguishing between these categories. If, as Smith assumes, the time-order error is negative, so that the first stimulus in a pair tends to be perceived as shorter and hence more /I/-like than the second, it would increase the discriminability of pairs of the type /I/-/i/ and /I/-/ɛ/ over the discriminability of the reverse order of these pairs. Since the effect would be mediated by the phonetic labels given to the stimuli, Smith's hypothesis fits well with the all-phonetic model of processing, and it predicts our results fairly well. There is a problem, however: two recent studies of duration discrimination using vowels comparable to our stimuli (Lehiste, 1976; Pisoni, 1976), have shown the time-order error to be positive, not negative. That is, the first stimulus in a pair tends to be perceived as longer than the second, probably due to the relatively short stimulus durations. In the light of this finding, Smith's hypothesis predicts just the opposite of her and our results. The hypothesis could be salvaged by assuming that, independently of the time-order error, there is a tendency to perceive /I/ as relatively longer than /i/ and /ɛ/, due to perceptual compensation when all vowels are of equal physical duration. If this is the case, a positive time-order error would tend to further increase the discrepancy in perceived duration when an /I/-like stimulus occurs first in a pair, thus enhancing discrimination. This is what

---

*Smith, M. R. (1976) An investigation of changes in categorical perception of vowels. Unpublished manuscript (Department of Linguistics, University of Connecticut).

we found. This explanation assumes, however, that the subjects base their discrimination responses on stimulus duration, not on phonetic labels. Therefore, it is not compatible with an all-phonetic model of vowel discrimination, nor does it fit into an all-auditory framework because of the mediating role of phonetic stimulus properties.

REFERENCES

Ades, A. E. (1977) Vowels, consonants, speech, and nonspeech. Psychological Review 84, 524-530.

Ainsworth, W. A. (1974) The influence of precursive sequences on the perception of synthesized vowels. Language and Speech 17, 103-109.

Bailey, P. J., A. Q. Summerfield and M. F. Dorman. (1977) On the identification of sine-wave analogues of certain speech sounds. Haskins Laboratories Status Report on Speech Research SR-51/52, 1-25.

Carney, A. E., G. P. Widin and N. F. Viemeister. (1977) Noncategorical perception of stop consonants differing in VOT. Journal of the Acoustical Society of America 62, 961-970.

Chistovich, L. A. and V. A. Kozhevnikov. (1970) Perception of Speech. In Theory and Methods of Research on Perception of Speech, ed. by L. A. Chistovich. (Washington, D.C.: Joint Publications Research Service, JPRS-50423), 1-101.

Crowder, R. G. (1971) The sound of vowels and consonants in immediate memory. Journal of Verbal Learning and Verbal Behavior 10, 587-596.

Crowder, R. G. (1973a) Representation of speech sounds in precategorical acoustic storage. Journal of Experimental Psychology 98, 14-24.

Crowder, R. G. (1973b) Precategorical acoustic storage for vowels of short and long duration. Perception and Psychophysics 13, 502-506.

Cutting, J. E., B. S. Rosner and C. F. Foard. (1976) Perceptual categories for musiclike sounds: Implications for theories of speech perception. Quarterly Journal of Experimental Psychology 28, 361-378.

Darwin, C. J. and A. D. Baddeley. (1974) Acoustic memory and the perception of speech. Cognitive Psychology 6, 41-60.

Eimas, P. D. (1963) The relation between identification and discrimination along speech and non-speech continua. Language and Speech 6, 206-217.

Finney, D. J. (1971) Probit Analysis, 3rd. edition. (Cambridge: University Press).

Fry, D. B., A. S. Abramson, P. D. Eimas and A. M. Liberman. (1962) The identification and discrimination of synthetic vowels. Language and Speech 5, 171-189.

Fujisaki, H. and T. Kawashima. (1969) On the modes and mechanisms of speech perception. Annual Report of the Engineering Research Institute, University of Tokyo 28, 67-73.

Fujisaki, H. and T. Kawashima. (1970) Some experiments on speech perception and a model for the perceptual mechanism. Annual Report of the Engineering Research Institute, University of Tokyo 29, 207-214.

Ganong, W. F. III. (1977) Selective adaptation and speech perception. Unpublished Ph.D. Dissertation, Massachusetts Institute of Technology.

Hall, L. L and S. E. Blumstein. (1977) The effect of vowel similarity and syllable length on acoustic memory. Perception and Psychophysics 22, 95-99.

Jamieson, D. G. and W. M. Petrusic. (1975) Presentation order effects in

duration discrimination. Perception and Psychophysics 17, 197-202.
Kanamori, Y., H. Kasuya, S. Arai and K. Kido. (1971) Effect of context on vowel perception. Seventh International Congress on Acoustics, Budapest. Paper 20 C4, pp. 37-40.
Lane, H. (1965) The motor theory of speech perception: A critical review. Psychological Review 72, 275-309.
Lehiste, I. (1976) Influence of fundamental frequency pattern on the perception of duration. Journal of Phonetics 4, 113-117.
Liberman, A. M., K. S. Harris, H. S. Hoffman and B. C. Griffith. (1957) The discrimination of speech sounds within and across phoneme boundaries. Journal of Experimental Psychology 54, 358-368.
Lindner, G. (1966) Veraenderung der Beurteilung synthetischer Vokale unter dem Einfluss des Sukzessivkontrastes. Zeitschrift fuer Phonetik, Sprachwissenschaft und Kommunikationsforschung 19, 287-307.
Macmillan, N. A., H. L. Kaplan and C. D. Creelman. (1977) The psychophysics of categorical perception. Psychological Review 84, 452-471.
Morse, P. A., J. E. Kass and R. Turkienicz. (1976) Selective adaptation of vowels. Perception and Psychophysics 19, 137-143.
Pastore, R. E., W. A. Ahroon, K. J. Baffuto, C. Friedman, J. S. Puleo and E. A. Fink. (1977) Common-factor model of categorical perception. Journal of Experimental Psychology: Human Perception and Performance 3, 686-696.
Peterson, G. E. and I. Lehiste. (1960) Duration of syllable nuclei in English. Journal of the Acoustical Society of America 32, 693-703.
Pisoni, D. B. (1971) On the nature of the categorical perception of speech sounds. Unpublished Ph.D. dissertation, University of Michigan.
Pisoni, D. B. (1973) Auditory and phonetic memory codes in the discrimination of consonants and vowels. Perception and Psychophysics 13, 253-260.
Pisoni, D. B. (1975) Auditory short-term memory and vowel perception. Memory and Cognition 3, 7-18.
Pisoni, D. B. (1976) Fundamental frequency and perceived vowel duration. Journal of the Acoustical Society of America 59 (Supplement no. 1), S39.
Pisoni, D. B. and J. H. Lazarus. (1974) Categorical and noncategorical modes of speech perception along the voicing continuum. Journal of the Acoustical Society of America 55, 328-333.
Pollack, I. and D. B. Pisoni. (1971) On the comparison between identification and discrimination tests in speech perception. Psychonomic Science 24, 299-300.
Pynn, C. T., L. D. Braida and N. I. Durlach. (1972) Intensity perception. III. Resolution in small-range identification. Journal of the Acoustical Society of America 51, 559-566.
Sachs, R. M. (1969) Vowel identification and discrimination in isolation vs. word context. Quarterly Progress Report, Research Laboratory of Electronics, Massachusetts Institute of Technology No. 93, 220-229.
Sachs, R. M. and K. W. Grant. (1976) Stimulus correlates in the perception of voice onset time (VOT): II. Discrimination of speech with high and low stimulus uncertainty. Journal of the Acoustical Society of America 60 (Suppl. no. 1), S91 (A).
Samuel, A. G. (1977) The effect of discrimination training on speech perception. Perception and Psychophysics 22, 321-330.
Stevens, K. N. (1968) On the relation between speech movements and speech perception. Zeitschrift fuer Phonetik, Sprachwissenschaft und Kommuni-

kationsforschung 21, 102-106.

Stevens, K. N., A. M. Liberman, M. Studdert-Kennedy and S. E. G. Ohman. (1969) Crosslanguage study of vowel perception. Language and Speech 12, 1-23.

Studdert-Kennedy, M., A. M. Liberman, K. S. Harris and F. S. Cooper. (1970) Motor theory of speech perception: A reply to Lane's critical review. Psychological Review 77, 234-249.

Thompson, C. L. and H. Hollien. (1970) Some contextual effects on the perception of synthetic vowels. Language and Speech 13, 1-13.

Tongue Position in Rounded and Unrounded Front Vowel Pairs

Lawrence J. Raphael[+], Fredericka Bell-Berti[++], René Collier[+++] and Thomas Baer

ABSTRACT

Traditional articulatory descriptions of front rounded and unrounded vowel pairs have assumed that tongue height is the same for the members of the pairs /i-y/, /e-ø/ and /ɛ-œ/. The electromyographic, articulatory synthetic, and acoustic investigations carried out in this study indicate that, in Dutch, the rounded member of the pairs /i-y/ and /e-ø/ was centralized. In the /ɛ-œ/ pair, however, the rounded vowel bears a different relationship to its unrounded counterpart.

INTRODUCTION

General phonetic descriptions of the articulation of vowels, as well as idealized or reference-grid schemes such as Daniel Jones' Cardinal Vowel System (Jones, 1940), portray pairs of rounded and unrounded vowels as having identical tongue heights (Figure 1). Thus vowel pairs such as /i-y/, /e-ø/, and /ɛ-œ/ are described as sharing the same degree of tongue height and tongue advancement, differing only in lip rounding. In those classification schemes employing a category of tongue tension, the members of each pair are said to share this feature as well. This sort of description has been rendered traditional through its repetition, with little or no modification in the writings of many phoneticians including Abercrombie (1967), O'Connor (1973), Smalley (1964) and Heffner (1964), among others. Delattre (1951), while noting that acoustic differences exist between rounded and unrounded front vowels, assigns the differences wholly to the frequency of the second formant and assumes that they are caused by differences in lip position and not tongue position. Viëtor (1921) provides a slightly different picture with regard to tongue advancement, with the rounded vowels being articulated

---

[+]Also Herbert H. Lehman College, City University of New York.

[++]Also Montclair State College (on leave).

[+++]Also University of Antwerp, Belgium and Institute for Perception Research, Eindhoven, The Netherlands.

Acknowledgment: We wish to thank Seiji Niimi for performing the electrode insertions and for making useful suggestions about the text. We also wish to thank Arthur S. Abramson with whom the idea for the 3-dimensional cardinal vowel diagram (Figure 1) originated. This work was supported, in part, by NINCDS Grant NS-13870; NICHD Grant HD-01994; NINCDS Grant NS-13617; NINCDS Fellowship Grant NS-05332; BRSG Grant RR-05596; and NFWO, Belgium.

[HASKINS LABORATORIES: Status Report on Speech Research SR-54 (1978)]

further back in the mouth than their unrounded counterparts, but here too tongue height for the relevant vowel pairs is equated.

Although the descriptions mentioned above originated in an attempt to provide a reference grid for vowels that were not intended to be language-specific, they are often applied literally to the description of languages having both rounded and unrounded front vowels. Assumed equivalence of tongue height is implicit, for example, in the instructions often given to English speakers learning a language such as French: "say /i/ and round your lips to produce the vowel of tu."

Dutch provides an example of a language that possesses front rounded and front unrounded vowels. The articulatory relationship between the two types of vowels received some attention a long time ago. In a 1928 study employing x-ray stills and palatography, Zwaardemaker and Eijkman described both /i/ and /y/ as closed vowels, and both /e/ and /ø/ as half-closed, although they reported a more advanced tongue position for the unrounded members of each pair. For the /ɛ - œ/ pair, a difference in overall mouth opening and presumably tongue height was reported between /ɛ/-half-open, and /œ/-closed. The authors also reported a difference in tongue tension and tongue advancement, with /ɛ/ being tense and front and /œ/ being lax and mid. For the members of each of the two vowel pairs that contrast primarily on the basis of lip rounding, the researchers found very similar averages and ranges for measurements of jaw opening.

Blanquaert's palatographic studies in the 1920's led him to conclude that although there may be some differences in tongue height and advancement between /i/ and /y/ and between /e/ and /ø/, "the main difference...must be sought in the position of the lips." He also noted that /ɛ/ and /œ/ are not related to each other in the same way as the members of the two other vowel pairs.

Wood (1975), in a cross-language x-ray study that did not include Dutch, reported that /y/ is generally articulated with a lower mandible position than /i/, but with equivalent tongue advancement.

Finally, in a formant analysis of Dutch vowels, Pols, Tromp, and Plomp (1973) found that the second- and third-formant frequencies are lower for the rounded front vowels than for their unrounded counterparts, as predicted by the acoustic theory of vowels (Stevens and House, 1955; Fant, 1960; Lindblom and Sundberg, 1971), if the members of each pair differ only in lip position. First-formant frequencies, however, are not always lower for the rounded member of each pair, as the acoustic theory would predict. These data are summarized in Table 1. We will discuss the articulatory implications of these acoustic measurements in connection with our own data below.

METHOD

In the present study, acoustic and electromyographic (EMG) analyses of vowels were performed on the speech of one native speaker of Dutch. The test utterances contained twelve Dutch vowels embedded in /ə pVp/ nonsense words, randomized in lists and repeated 24 to 30 times each. Six of the twelve vowels constituted the three front rounded-unrounded vowel pairs that were the object of this investigation. Examples of the test utterances are /əpip/,

/əpyp/, /əpap/, and /əpup/. Hooked-wire electrodes were inserted into the genioglossus (anterior fibers), mylohyoid and anterior belly of the digastric muscles using standard procedures that are described elsewhere (Hirose, 1971; Raphael and Bell-Berti, 1975). EMG potentials were also recorded from the orbicularis oris muscle, which is active in rounding the lips. These data are not discussed below, since, as we might expect, this muscle showed considerably more activity for the rounded than for the unrounded members of the pairs /i-y/, /e-ø/ and /ɛ-œ/, and this activity is quite similar for the rounded vowels /y,ø,œ/. The EMG potentials were then rectified, integrated and computer-averaged. The EMG signals and functions derived from them were aligned with reference to the onset of the voicing of the stressed vowel of each utterance.

The acoustic analyses were performed using a digital waveform and spectral analysis system. Formant frequencies for 9 to 12 repetitions of each stressed vowel were determined at a point approximately equidistant from the surrounding consonant closures. That is, the measurements were made for that portion of each vowel that most closely approximated a steady state.

RESULTS

Acoustic Analyses

We will confine our discussion here to the front pairs of rounded and unrounded vowels that are found in Dutch: /i-y/, /e-ø/, and /ɛ-œ/.

The formant frequency values resulting from the analysis of our subject's speech (Table 1) generally agree with the data of Pols, Tromp, and Plomp (1973) as to the relationships between these vowels, as can be seen in Table 1 where values are given for the first three formants.[1] The F_1 and F_2 data for our subject are also presented in Figure 2. All of these data agree with the predictions, presented above, of the effect of rounding on F_2 and F_3, since lip rounding effectively lengthens the vocal tract, lowering the second- and third-formant frequencies. On the other hand, these data do not always agree with the prediction that the effect of rounding on first-formant frequency is to lower it.

The articulatory implications of these data are that the vocal-tract differences between members of pairs cannot be due to lip rounding alone. We assume, in the following discussion, that relative tongue position can be inferred from an F_1-F_2 frequency plot for vowels having the same lip position. Since lip rounding is expected to lower both F_1 and F_2, and F_1 is higher for the rounded members of the /i-y/ and /e-ø/ pairs, the tongue must be lower for the rounded members of these pairs. However, we cannot describe these differences in tongue position on the basis of the acoustic data alone.

---

[1]The differences between our data and those of Pols, Tromp, and Plomp (1973) may have two different causes. First, our data are measurements of one speaker's productions of the target vowels in a /əpVp/ frame, while the Pols, Tromp, and Plomp data are averages of 50 speakers' productions of the target vowels in a /hVt/ frame. Second, our speaker's dialect is Southern Dutch, while the speakers in the Pols, Tromp, and Plomp study spoke Northern Dutch.

TABLE 1: Averaged formant frequencies for one native Dutch speaker. The number of utterances averaged for each vowel pair ranged from 9 to 12.

	/i/	/y/	/e/	/ø/	/ɛ/	/œ/
F_1	242	277	341	375	538	382
F_2	2006	1691	1956	1530	1508	1400
F_3	2902	2111	2669	2229	2377	2238

Averaged formant frequencies for 50 native Dutch speakers. From Pols, Tromp, and Plomp, 1973.

F_1	294	305	407	443	583	438
F_2	2208	1730	2017	1497	1725	1498
F_3	2766	2208	2553	2260	2471	2354

The higher first-formant values for the rounded members of the /i-y/ and /e-ø/ pairs indicate centralization on the vertical axis. In the case of the /ɛ- œ/ pair, the situation is less clear. The lower first-formant value for /œ / suggests that this vowel is no closer to the tongue height of a mid-central vowel than is the unrounded /ɛ/, but rather, that the tongue height is closer to that of the half-closed vowels /e/ and /ø/. In general, however, centralization of tongue position, as inferred from formant-frequency measurements, seems to mark the front rounded vowels and to distinguish them from the front unrounded vowels.

EMG Analyses

Before presenting the EMG data it will be necessary to make explicit certain assumptions underlying their interpretation. The first assumption is that if, in a constant framework, the EMG potentials recorded from the tongue muscles are different that tongue position and, hence, vowel quality, will be different.[2]

The second assumption (and those that follow it) concerns muscle function: the genioglossus is the only muscle that contributes significantly to tongue advancement (fronting) for the front vowels being considered here. Further, it is assumed that virtually all tongue fronting gestures for low vowels can be accounted for by relatively moderate amounts of genioglossus activity. We are thus led to a third assumption, which is that genioglossus activity that exceeds levels needed for near-maximum tongue fronting for low vowels contributes primarily to the raising and bunching of the tongue as well as to further tongue advancement. The second and third assumptions taken together suggest that with the tongue body low in the mouth, relatively little muscular contraction is needed to push the tongue as far forward as it will go. More contraction, assuming the tongue tip to be bent down and resting behind the lower teeth, will cause the center of the tongue to rise (bunch) toward the post-alveolar area of the palate and to be increasingly more advanced as it rises to its highest and most fronted position. We suggest, then, that the direction of the rise of the high point of the tongue attributable to genioglossus contraction is oblique, along an anterior-superior line, and that the strength of genioglossus contraction will be roughly proportional to tongue height, for front vowels (Smith[3]; Perkell[4]; Raphael and Bell-Berti, 1975; and Kakita, 1976).

---

[2]This assumption concerns the nature of EMG activity itself and its relationship to articulator movement. Although the relationship between EMG activity and muscle tension is not linear, we assume that within a constant framework the relationship is monotonic, and therefore, that relatively stronger EMG potentials result in relatively greater muscle tension and, therefore, greater articulator displacement.

[3]Smith, T. St. J. (1970) A Phonetic Study of the Function of the Extrinsic Tongue Muscles. Unpublished doctoral dissertation, U.C.L.A.

[4]Perkell, J. (1974) A Physiologically-Oriented Model of Tongue Activity in Speech Production. Unpublished doctoral dissertation, M.I.T.

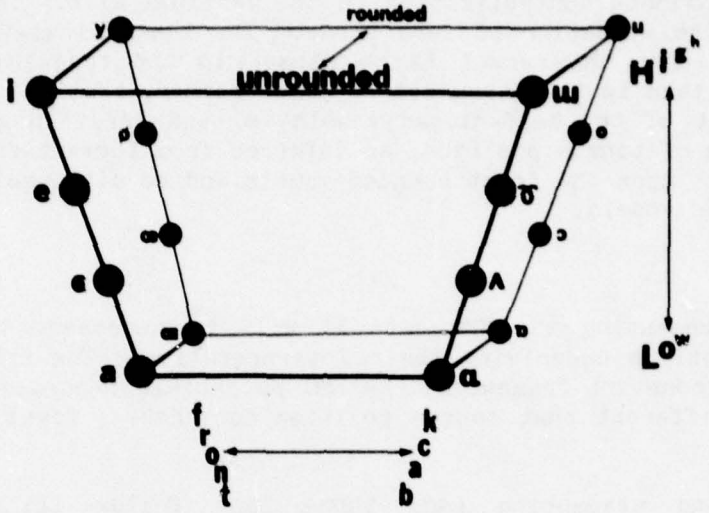


Figure 1: Relative tongue positions for selected cardinal vowels.

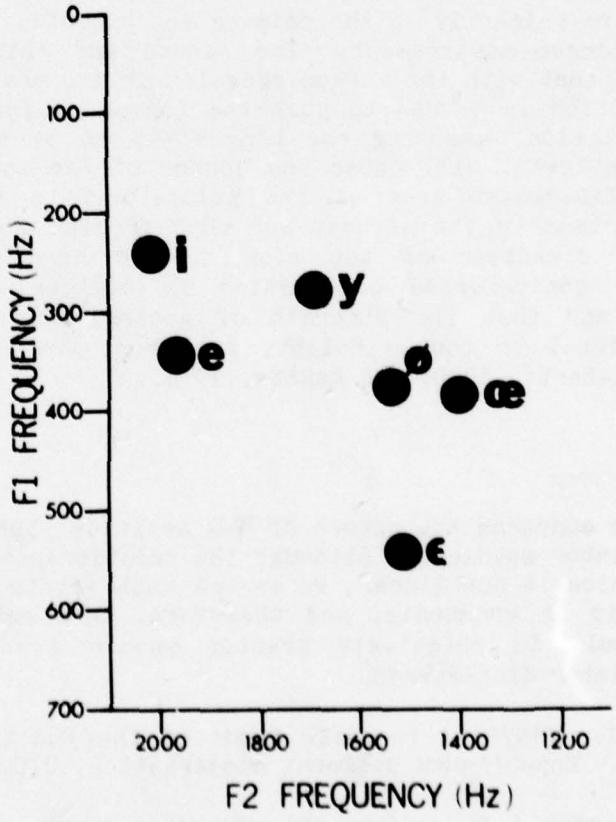


Figure 2: Frequency of the first formant versus frequency of the second formant for rounded and unrounded front vowels for a speaker of Dutch.

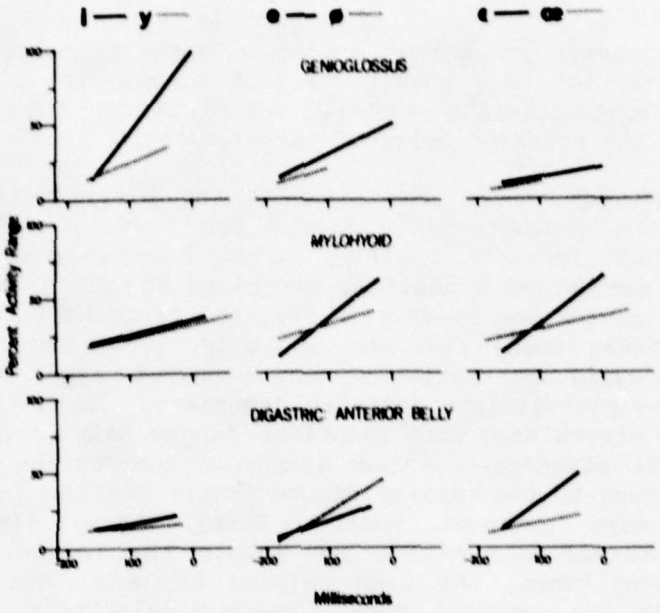


Figure 3: Normalized EMG data, schematized as percent of maximum range of activity, from onset of activity to peak of activity, for each muscle.

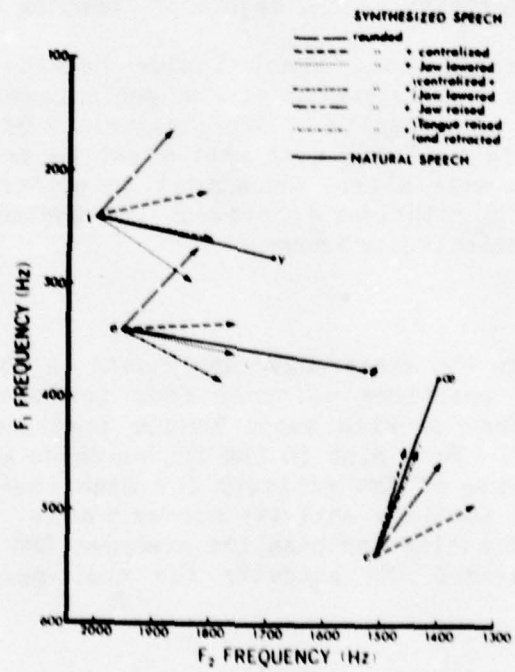


Figure 4: Frequency of the first formant versus frequency of the second formant for natural and synthetic rounded and unrounded front vowel pairs. Solid lines connect natural-speech values for each pair. The arrows indicate the direction of formant-frequency change resulting from articulatory adjustments, as described in the text.

Our fourth assumption is that mylohyoid activity raises the tongue body through the application of a nearly vertical force. The effect of mylohyoid contraction on tongue height varies, depending on the activity of the genioglossus and the anterior belly of the digastric.

Previous research throws some light on the possible interaction of the mylohyoid with the genioglossus. First, for vowels, both muscles display maximum contraction for /i/ (Harris, 1971; Faaborg-Anderson and Vennard, 1964). Second, genioglossus activity decreases for the front vowel series as the tongue is retracted and lowered (Smith, see footnote 3; Raphael and Bell-Berti, 1975). Given these findings, we would expect that, all other things being equal, mylohyoid contraction becomes proportionately more important for tongue raising as genioglossus activity decreases. Thus, for example, if two front vowels are articulated with identical tongue heights but with different degrees of tongue advancement, then mylohyoid contraction should contribute proportionately more to the raising of the tongue for the less advanced vowel than for the more advanced vowel. This follows simply because the genioglossus simultaneously raises and fronts the tongue: thus, the less fronting it accomplishes, the less raising as well, and so the mylohyoid contraction, with its vertical force, becomes relatively more important in maintaining tongue height.

Our final assumption is that the effect of activity of the anterior belly of the digastric is to lower the jaw and, in lowering the jaw, it counteracts the activity of the mylohyoid, and, to a lesser extent, the activity of the genioglossus. Jaw lowering increases, for the vowels considered here, only as articulation occurs progressively further back in the mouth, and as we have seen, tongue height for front vowels depends proportionally more on mylohyoid than on genioglossus activity as the degree of fronting decreases.

At a first approximation, then, tongue height for front vowels is determined by the combined activity of the genioglossus and mylohyoid, less the activity of the anterior belly of the digastric.[5] Of course, this is not a formula for tongue height of the sort that might be proposed if the anterior belly of the digastric were a true antagonist to either the mylohyoid or the genioglossus, or if the relationship between the various muscular forces and EMG measures were quantitatively known.

The EMG Data

Let us turn now to the electromyographic data to see if they confirm the relative front vowel positions inferred from the acoustic data, and what additional insights they provide about tongue position. The EMG data are displayed in Figure 3. Each plot in the figure shows a schematized representation of the time-course of EMG activity for each vowel as a percent of the overall range of each muscle's activity across vowels. The range of activity was determined by subtracting the baseline averaged EMG activity for a muscle from the maximum averaged EMG activity for that muscle for any utterance

---

[5]Of course, specifying tongue height completely requires additional information, including, but not limited to, the activity of the internal pterygoid, the muscle that raises the jaw in speech.

(including the vowels not discussed in this paper). In Figure 3, the minimum points were derived by dividing the averaged EMG activity (after subtracting baseline activity) at the beginning of the vowel gesture by the range of activity. Similarly, the maximum points were derived by dividing the maximum averaged EMG activity for the vowel (after subtracting baseline activity) by the range of activity. Activity is plotted as a line from the point of onset to the point of peak vowel activity.

Tongue movements toward a vowel may begin before voice onset for the vowel and, in addition, there is a time delay between EMG activity and its movement consequences. Thus EMG activity commences some time before the vowel is heard. The moment of initiation of voicing for a given vowel is marked by the 0 point at the right of the abscissa in each graph. It will be recalled that this is the point of alignment of the EMG signals for all the tokens of a given utterance type. We shall consider each pair of vowels in turn.

/i/ vs. /y/: The moderate degree of genioglossus activity for /y/ serves to advance the high point of the tongue and to raise it, but not to the extent that it is raised and advanced by the comparatively more vigorous contraction for /i/. For this vowel pair the activity of the other muscles is either at a relatively low level (that is, anterior belly of the digastric) or the activity is essentially the same from one vowel to the other (that is, mylohyoid). Thus, the tongue is apparently somewhat higher and more advanced for the unrounded member of this vowel pair.

/e/ vs. /ø/: The differences in genioglossus, mylohyoid, and anterior belly of the digastric activity for this vowel pair indicate that the tongue is both higher and more advanced for /e/ than for /ø/. Both the genioglossus and the mylohyoid are more active for /e/, producing both greater tongue height and advancement, while the anterior belly of the digastric is more active for /ø/, implying more jaw lowering for the rounded vowel. Taken together, these EMG data indicate that the tongue is less advanced and lower for the rounded vowel of the pair.

/ɛ/ vs. /œ/: Genioglossus activity is at a relatively low level for both vowels of this pair. Only /ø/, among the other vowels, displays the same low level of genioglossus activity found for both members of this pair. However, the genioglossus activity, persisting for substantially longer for /ɛ/ than for /œ/, seems to indicate slightly more tongue advancement for the unrounded vowel than for its rounded counterpart. Both the mylohyoid and the anterior belly of the digastric are considerably more active for /ɛ/ than for /œ/, indicating that both tongue raising and jaw lowering are greater for the unrounded vowel of this pair: it is impossible, however, to determine which is more effective. Obviously, though, we cannot describe tongue height differences from the EMG data alone.

The acoustic analysis suggested that /ɛ/ and /œ/ differ from the other vowel pairs in that the rounded vowel appears to have a higher tongue position than the unrounded vowel. This height difference may be attributed to the interaction of the activity of the mylohyoid and the anterior belly of the digastric muscles. Although there is more mylohyoid activity for /ɛ/ than for /œ/, this activity is counteracted by the contraction of the anterior belly of the digastric for the /ɛ/. This finding closely parallels Zwaardemaker and Eijkman's 1928 description of /ɛ/ as half-open and of /œ/ as closed, although

we would prefer half-closed as a descriptor for /œ/ on the basis of the EMG and acoustic data. On the other hand, tongue height may be essentially the same for these vowels, with lip rounding entirely accounting for the lower first-formant frequency of the rounded member of the pair.

Analysis by Articulatory Synthesis

The conclusions drawn from the acoustic and EMG data were tested using an articulatory synthesizer (Mermelstein, 1973; Cooper, Mermelstein, and Nye, 1977). An exemplar of each of the unrounded vowels was produced, and the first three formant frequencies were recorded. The behavior of the model with respect to the first three formants was then investigated as a function of: lip rounding alone, and lip rounding together with jaw and tongue position adjustments. Since the changes in the frequencies of both the second- and third-formants were always in the same direction, we have plotted only the second-formant values in Figure 4.

As expected, lip rounding alone always lowered all three formant frequencies. The increase of the first-formant frequency for the rounded members of the /i-y/ and /e-ø/ pairs could be produced with either a lowered jaw position or a lowered jaw position together with centralization of the tongue body. For the /ɛ-œ/ pair, changes in the direction of those observed in our native speaker could be produced with lip rounding and a combination of jaw raising and tongue raising and retracting. Thus, our hypothesis about articulatory movements, based on the acoustic and EMG data, appear to be confirmed.

DISCUSSION

Both the electromyographic (especially genioglossus) and acoustic data of this study indicate that the front unrounded vowels /i/ and /e/ are articulated with the high point of the tongue in a higher and more advanced position than their rounded counterparts /y/ and /ø/. (This conclusion was supported by the results of the articulatory synthesis experiment reported above.) To put it another way, we might say that the rounded vowels are marked by centralization of the high point of the tongue in relation to their unrounded counterparts and not only differences in tongue height, as suggested by Wood (1975). Thus we can see from the EMG data that the lowered second- and third-formant frequencies result, in part, from differences in tongue position, and not only from the increased cavity length attributable to lip rounding. It is of note that these differences in tongue position are caused in different ways: in the /i-y/ case the difference is primarily in genioglossus activity, while in the /e-ø/ case the different positions appear to be caused by differences in mylohyoid and anterior belly of the digastric activity as well as genioglossus activity. We might add that this point cannot be derived from the analysis of the acoustic signal alone.

The data further indicate that there is a qualitative difference between the two vowel pairs /i-y/ and /e-ø/ and the third pair /ɛ-œ/. It seems clear that /ɛ/ and /œ/ do not constitute a rounded-unrounded vowel pair in the same sense that /i-y/ and /e-ø/ do. In this we find ourselves in substantial agreement with other investigators (Zwaardemaker and Eijkman, 1928; Blanquaert, 1969; Pols, Tromp and Plomp, 1973).

We might also note, in passing, that the dimension of vowel height is implemented differently from the unrounded to the rounded series of vowels. For the unrounded /i-e-ɛ/ series we find: (1) decreasing genioglossus activity; (2) increasing anterior belly of the digastric activity; and (3) an increase in mylohyoid activity from /i/ to /e/, but not from /e/ to /ɛ/. For the rounded /y-ø-œ/ series we find: (1) a more subtle gradation in genioglossus activity; (2) an increase in anterior belly of the digastric activity from /y/ to /ø/; but a decrease in activity from /ø/ to /œ/; and (3) very little variation in mylohyoid activity.

In addition, apart from the question of the relative tongue position in the pairs considered here, we might note that the apparent difference in tongue height between /ɛ/ and /œ/ confounds traditional descriptions of their positions: the distance between /e/ and /ɛ/ is much greater than the distance between /ø/ and /œ/.

Further investigation is indicated to determine if lowering and centralization of tongue position is a general property of the so-called front rounded vowels in relation to their unrounded counterparts in languages other than Dutch (and in speakers other than our subject). One might also wish to discover whether some other language such as Danish or Turkish possesses a third pair of front vowels the members of which are related to each other as /i/ and /e/ are related to /y/ and /ø/, respectively, in Dutch.

As a final note, we find it interesting that lowering and centralization characterize rounded (front) vowels in their opposition to the corresponding unrounded ones, and, at the same time, characterize "lax" vowels in their opposition to their "tense" counterparts (Ladefoged, 1975).

REFERENCES

Abercrombie, D. (1967) Elements of General Phonetics. (Chicago: Aldine).

Blanquaert, E. (1969) Praktische Uitspraakleer van de Nederlandse Taal, 8th edition. (Antwerp: DeSikkel).

Cooper, F. S., P. Mermelstein and P. W. Nye. (1977) Speech synthesis as a tool for the study of speech production. In Dynamic Aspects of Speech Production, ed. by M. Sawashima and F. S. Cooper. (Tokyo: University of Tokyo Press).

Delattre, P. (1951) The physiological interpretation of sound spectrograms. Modern Language Association of America, vol. LXVI, no. 5, 864-875.

Faaborg-Andersen, K. and W. Vennard. (1964) Electromyography of extrinsic laryngeal muscles during phonation of different vowels. Annals of Otology 73, 248-254.

Fant, C. G. M. (1960) Acoustic Theory of Speech Production. ('s Gravenhage: Mouton).

Harris, K. S. (1971) Action of the extrinsic musculature in the control of tongue position: preliminary report. Haskins Laboratories Status Report on Speech Research SR-25/26, 87-96.

Heffner, R.-M. S. (1964) General Phonetics. (Madison, Wisc.: University of Wisconsin Press).

Hirose, H. (1971) Electromyography of the articulatory muscles: current instrumentation and technique. Haskins Laboratories Status Report on Speech Research SR-25/26, 73-86.

Jones, D. (1940) An Outline of English Phonetics, 6th edition. (New York:

E. P. Dutton).
Kakita, K. (1976) Activity of the genioglossus muscle during speech production: An electromyographic study. Unpublished D. M. S. dissertation, University of Tokyo.
Ladefoged, P. (1975) Course in Phonetics. (New York: Harcourt Brace Jovanovich).
Lindblom, B. E. F. and J. E. F. Sundberg. (1971) Acoustical consequences of lip, tongue, jaw, and larynx movement. Journal of the Acoustical Society of America 50, 1166-1179.
Mermelstein, P. (1973) Articulatory model for the study of speech production. Journal of the Acoustical Society of America 53, 1070-1082.
O'Connor, J. D. (1973) Phonetics. (Baltimore: Penguin Books).
Pols, L. C. W., H. R. C. Tromp and R. Plomp. (1973) Frequency analysis of Dutch vowels from 50 male speakers. Journal of the Acoustical Society of America 53, 1093-1101.
Raphael, L. J. and F. Bell-Berti. (1975) Tongue musculature and the feature of tension in English vowels. Phonetica 32, 61-73.
Smalley, W. A. (1964) Manual of Articulatory Phonetics. (Tarrytown, N.Y.: Practical Anthropology).
Stevens, K. N. and A. S. House. (1955) Development of a quantitative description of vowel articulation. Journal of the Acoustical Society of America 27, 484-493.
Viëtor, W. (1921) Elemente der Phonetik des Deutschen, Englischen und Franzosischen. (Leipzig: Reisland).
Wood, S. (1975) The weakness of the tongue-arching model of vowel articulation. Phonetics Laboratory, Lund University, Working Papers 11, 55-108.
Zwaardemaker, H. and L. P. H. Eijkman. (1928) Leerboek der Phonetiek. (Haarlem: DeErven F. Bohn).

The Reading Behavior of Dyslexics: Is There a Distinctive Pattern?*

Donald Shankweiler+ and Isabelle Y. Liberman+

ABSTRACT

Few positive signs have been proposed for the differential diagnosis of dyslexia from among the wider class of backward readers. Reversals in reading letters and words are frequently cited and have been regarded as symptomatic of a perceptual deficit underlying dyslexia. Our findings do not support the view that a subclass of dyslexics can be differentiated from other poor readers on the basis of a high frequency of reversal errors, but some children clinically diagnosed as dyslexic show orientational and directional biases that are absent in most poor readers. Moreover, the difficulties manifested in their common error pattern are, in the main, language related and are not correctly attributed to anomalies of visual perception. Consideration is given to the manner in which linguistic factors may influence the reading behavior of dyslexics and other poor readers.

INTRODUCTION

Research on reading disability has produced rather little that is of use in diagnosis and treatment despite considerable expenditure of individual effort and public money. The low yield of research on reading problems no doubt has a number of causes. A major one, in our view, is that the search for causes of reading disability has proceeded independently of investigation into the foundations for reading acquisition in the normal child. For the past several years, the reading research group at Haskins Laboratories has been asking questions about how learning to read builds upon the earlier speech acquisitions of the child. We think we have made progress in understanding the relationships between reading and spoken language and that our findings throw some light on the causes of reading disability, including

---

*This paper was presented at The Fourth Biennial Congress of The International Society for The Study of Behavioral Development, held at The University of Pavia (Italy), 19-23 September, 1977. It will appear in the congress proceedings, titled Cognitive Aspects, vol. 2, ed. by O. Andreani (in Italian). It will also be published in The Bulletin of the Orton Society.

+Also University of Connecticut.

Acknowledgment: Much of the authors' research on reading acquisition was supported by a grant to Haskins Laboratories from the National Institute of Child Health and Human Development (Grant HD 01994).

[HASKINS LABORATORIES: Status Report on Speech Research SR-54 (1978)]

the special case of dyslexia.

A necessary condition for considering a child "dyslexic" is the existence of a significant disparity between the child's actual measured performance and the level of reading performance that might reasonably be expected in view of his intelligence and his educational opportunities. This would serve to distinguish the dyslexic child from the child whose educational attainments are uniformly low. Of course, more has usually been implied in the use of the term. The designation "dyslexic" carries an implication about the causes of reading failure. It assumes a constitutional inadequacy that blocks the efficient acquisition of reading while not resulting in general intellectual retardation (Benton, 1975).

PERCEPTION OF LETTER ORIENTATION AND LETTER SEQUENCE IN DYSLEXIA

Much effort has gone into the search for pathognomic signs of dyslexia--that is to say, qualitative features that distinguish a dyslexic child from one who is merely backward in reading. The tendency of young children to confuse letters of similar shape that differ in orientation (such as b,d,p,q) is well known. Reversal of the direction of letter sequences (such as reading "was" for saw) is another phenomenon that is frequently cited and usually considered to be intrinsically related to orientation reversal. Both types of reversals have been viewed as symptoms of a disturbance in the visual directional scan of print in children with reading disability. One early student of reading problems, S. T. Orton (1937), believed that reversals are of diagnostic significance in dyslexia; indeed, so convinced was he of their centrality that he invented the name "strephosymbolia" to designate the condition of specific childhood reading disability. Reversal phenomena, in Orton's view, are a manifestation of poorly established left cerebral hemisphere dominance for speech.

In view of the continuing influence of Orton's views and the persisting unanswered questions about the role of reversals in dyslexia, we were prompted to ask whether children diagnosed as dyslexic exhibit a distinctive pattern of misreadings and, in particular, to discover whether they reverse letters and words more frequently than other poor readers whose backwardness stems from diverse causes.

The pattern of errors in reading isolated words and nonsense syllables was studied in two groups of children, aged 8 and 10 years, all within the normal range of intelligence. One group (see Fischer, Liberman and Shankweiler, in press) consisted of children diagnosed as "dyslexic" by the staff of the Kennedy Institute of the Johns Hopkins Hospital, Baltimore. The other group included all the children in the second year of a Connecticut elementary school who fell into the lowest third on a standard test of reading achievement (Liberman, Shankweiler, Orlando, Harris, and Bell-Berti, 1971). Although the dyslexic children were somewhat pooorer in word recognition than the backward readers selected purely on psychometric grounds, the groups did not differ significantly in the incidence of reversal errors. When each error type (see Figure 1) was tabulated as a proportion of the total opportunities for an error of that type to occur, it was found that reversals of letter sequence (RS) occurred with an incidence of 8 percent among the dyslexic group and 6 percent among the school group. Similarly, reversals of letter

orientation (RO) occurred with an incidence of 12 percent and 13 percent, respectively. Thus the groups appeared nearly alike in their tendency to make each type of reversal error. Moreover, for both groups, reversals represented only a small proportion of the total number of reading errors. Vowel errors and errors on nonreversible consonants provided the bulk of the misreadings for both the dyslexics and backward readers. The low incidence of reversal errors in both groups should surely raise questions regarding Orton's belief that reversals are the hallmark of dyslexia.

The performances of the dyslexic and school groups did differ, however, in regard to certain spatial characteristics of letter reversal errors. The dyslexics showed a 2:1 excess of horizontal over vertical letter transformations (for example, b was confused with d more often than with p). There was, moreover, an asymmetry in the direction of reversals; namely, there was a bias to read reversible letters from right to left (b going to d instead of d going to b). Neither of these features was characteristic of the school group of poor readers. For them, vertical reversals occurred as frequently as reversals in the horizontal plane and horizontal reversals did not show a directional bias. The absence of this directional bias in the errors made by the school group, together with its presence in the dyslexic group, suggests that reversible letters may present a special obstacle to some dyslexic children, lending a measure of support to Orton's claims. Further work is needed to discover whether children who show a consistent directional bias in their reversal errors can be distinguished in other ways from the larger group who do not.

On balance, however, we were more impressed by the similarities in the results for the two groups than by the differences. It may be seen from Figure 1 that for both groups, more errors occurred on nonreversible consonants than on reversible ones, and vowels elicited the highest rate of error. Errors on the final consonant of consonant-vowel-consonant (CVC) syllables were about double those on the initial consonant, while errors on the medial vowel exceeded those on consonants in both initial and final position (Fischer, Liberman and Shankweiler, in press). In addition, the same error pattern occurred in reading nonsense syllables as in reading words, though, of course, the actual frequencies differed.

It appears that the common error pattern is determined largely by the phonetic and orthographic structure of words. One aspect of the common pattern, that vowels elicit more errors than consonants, is well documented in earlier work (Monroe, 1932; Weber, 1970; Shankweiler and Liberman, 1976) and remains true when the position of the vowel is varied in the word. The difference in consonant and vowel error pattern suggested that they might have different causes. This idea was supported by the results of an analysis that took account of the phonetic relationships between consonants and vowels as written and the sounds substituted for them when misread.

If reading involves the conversion of the graphic shapes of print into a speech-based internal representation, then we might expect misreadings of a word to bear a systematic phonetic relationship to the target word. This expectation was clearly borne out in data on consonant errors that we obtained with a random selection of school children (Fowler, Liberman and Shankweiler, 1977). Misreadings of consonants were tabulated according to the number of

phonetic features shared between the phonological segments of the target word and the phonological segments of the word as read. Consonant substitutions were found to bear a close phonetic relationship to the target word, differing most often in only one of three phonetic feature values (voicing, place of production, manner of production). Vowel errors, in contrast, were not systematically related to the phonetic features of the vowel as written (tenseness, tongue advancement, tongue height and diphthongization). Thus, the dimensional analysis, so successful in rationalizing the substitutions among the consonants, does not enable us to understand the vowel errors. Here, phonetic contributions to the error pattern are presumably obscured by some other more powerful source.

It is surely significant that the opposition between consonant and vowel, which occurs universally in speech, should be manifested in the pattern of reading errors. We suspect that the difference in error pattern of vowels and consonants is related to the different functional roles they have in English phonology. Vowels are the more fluid and variable of the two classes of phones, more subject to phonetic drift over time. This relatively greater variability of vowels may account, in part, for their more complex representation in the spelling system of English, particularly for the fact that there tend to be many spellings for each vowel and more nearly one-to-one spelling-to-sound relationships for consonants. Either factor would account for the higher rate of misreading of vowels than consonants. At present we are attempting to discover, by cross-language comparisons, whether the preponderance of vowel errors over consonant errors disappears in reading a language (Serbo-Croatian) in which the representation of the vowels is more nearly phonetic than is the case in the orthography of English.

In any event, it is clear to us that the pattern of reading errors in the beginning reader, whether dyslexic or not, has to be understood chiefly in linguistic terms, not in terms of the properties of letters as shapes. Taken as optical shapes, the set of letters representing consonants is not marked in any distinctive way from the set representing the vowels. Therefore, it is almost inconceivable that the differences between consonants and vowels in frequency of misreadings, in distribution of errors within the syllable, and in the nature of the errors could reflect misclassification based on visual characteristics.

READING DISABILITY AND LANGUAGE DEFICITS

Our research on language deficits and reading disability was carried out with school children who had been selected for backwardness in reading behavior. How many would have met the restrictive criteria for selection of the Institute Group is open to question. However, we believe that the approach we adopted has value for future studies of dyslexia, narrowly defined.

We have shown that poor readers frequently show subtle deficits in language development that manifest themselves not as clinically detectable gross difficulties in speaking and understanding, but instead as failures to achieve awareness of the phonetic structure of language so necessary for effective use of an alphabet (Liberman, Shankweiler, Fischer and Carter, 1974;

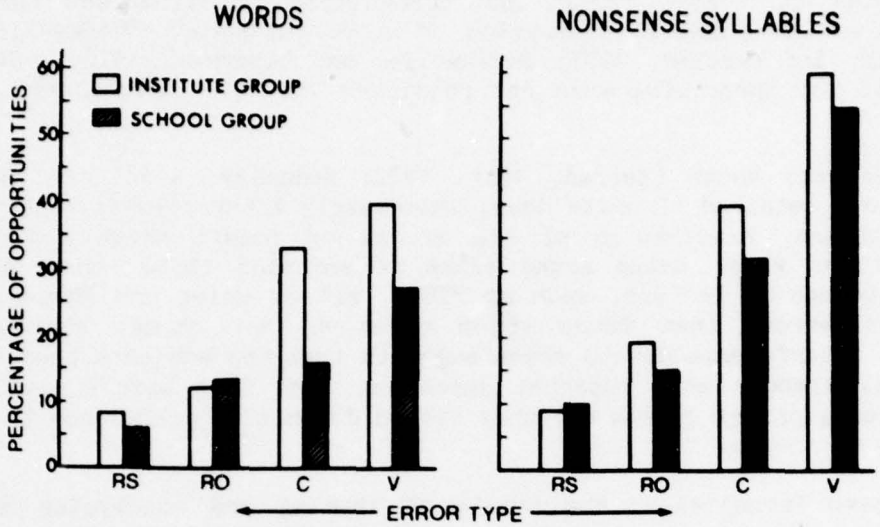


Figure 1: Errors in relation to opportunities by the Institute group and the School group for monosyllabic words and nonsense syllables. RS = reversed sequence; RO = reversed letter orientation; C = other consonant error; V = vowel error.

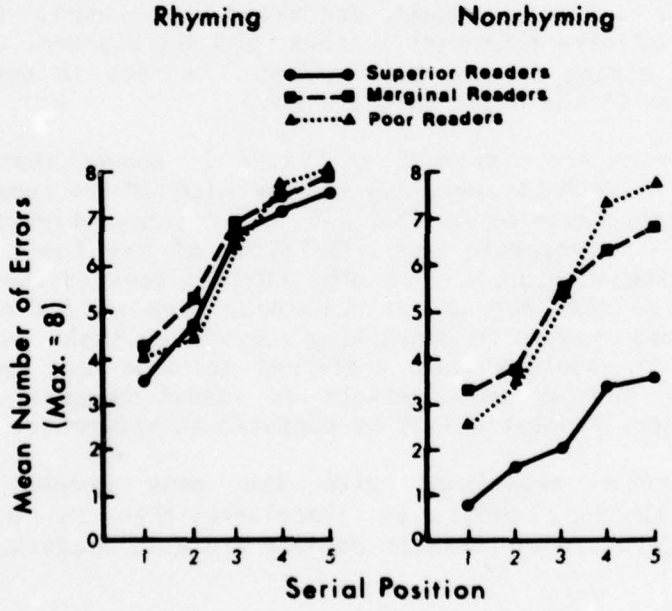


Figure 2: Errors in recall of rhyming and nonrhyming letter strings plotted by serial position for three groups of children who differed in reading achievement (from Liberman et al., 1977).

Zifcak[1]). We have also found a language deficit of another kind to be characteristic of children who are backward to reading: they tend to show deficiencies in verbal recall and recognition in situations that require retention of the phonetic properties of words (Liberman, Shankweiler, Liberman, Fowler and Fischer, 1977; Shankweiler and Liberman, 1976). This latter finding is not surprising when one considers the role of phonetic coding in reading.

It is well known (Conrad, 1964, 1972; Baddeley, 1968) that when adult subjects are required to write down, immediately after presentation, a list of random letters, examination of the errors of recall shows a tendency to confuse items whose names sound alike rather than those whose shapes look alike. Strings of letters, such as PTBDG, all of which are rhymes, generate many more errors than those whose names do not rhyme, such as WLRZY. Selective interference due to rhyme suggests that the subjects tend to convert the visual symbols into internal speech at some time before recall, though there is no a priori reason why they should do this in preference to retaining the items as shapes.

We have investigated the recall of rhyming and nonrhyming strings of letters as a way of examining the possibility that children who read well may differ from those who read badly in the process by which they transform visual material into speech. We considered that, given the task of serial recall of random strings of rhyming items, children who are good readers might be more apt to transform the stimuli into speech and might, therefore, be more susceptible to phonetic interference than backward readers. Three groups of school children aged 8 to 10 were the subjects in this study (Liberman et al., 1977). Roughly matched in IQ, they differed in level of reading attainment (as assessed by a standard test, the Wide Range Achievement Test). They were designated as superior, mildly backward and severely backward. These subjects were shown strings of five consonant letters, briefly exposed, and were asked to write down each string in the order given. Letters in one-half of the strings rhymed; those in the other half did not.

The results, which are displayed in Figure 2, showed that rhyme had a strong penal effect on recall, but only in the case of the superior readers. In contrast, the performance of the two groups of backward readers was much less influenced by the phonetic characteristics of the items (that is, by whether or not the items rhymed). The most obvious possibility seemed to be that the backward readers, for whatever reason, were not as vigorous or as consistent as the good readers in converting the visual input into a phonetic representation. Thus, because they preferred to code the letters not as speech items but in another way (perhaps as visual shapes), their recall performance showed less susceptibility to phonetic interference.

Through a further experiment with the same groups of children (Shankweiler and Liberman, 1976), we discovered that the differences we obtained in the performance of backward readers and good readers do not depend

---

[1]Zifcak, M. (1977) Phonological awareness and reading acquisition in first grade children. Unpublished doctoral dissertation, University of Connecticut.

on visual presentation; the findings were essentially reproduced when the stimuli were presented by ear. To carry out an auditory serial recall experiment necessitated a further condition in the visual mode in which the letters in each group were presented successively one by one rather than as a simultaneous array. The two new conditons were thus precise analogues that differed only in modality. All three studies, the two with visual presentation and the one with auditory, gave strikingly similar results with regard to the question of interest. In each, the backward readers were relatively little influenced by the phonetic characteristics of the items, whereas the superior readers were greatly affected by that variable.

The new results forced us to revise our opinion about the nature of the problem the backward readers were having. No longer could we see the transformation of visual stimuli into a phonetic representation as the crux of the problem, since differences of the same order of magnitude occurred when the stimuli were spoken and presented to the ear. Apparently, the internal representation of the group of stimuli persists longer or is more accessible in the good reading subjects, regardless of whether stimulus presentation is to the eye or to the ear.

How, then, are we to regard the backward readers' difficulty? One might ask whether backward readers are generally impaired on all memory tasks, whatever their nature. In reply we can state that it has not been found that backward readers are consistently impaired on memory tasks, other than those involving linguistic material or others on which speech coding may readily occur (see Vellutino, 1977 for a discussion of this question).

Another possibility we must consider is that the underlying difficulty of the backward readers is in recall of the temporal order of the elements of an auditory or visual pattern. Two facts weigh heavily against this interpretation. First, let us return to our experiment on recall of rhyming and nonrhyming letter strings. This was, of course, a serial recall task in that the subjects had to recall the left-right order or the temporal order of the items in making their responses. We (Shankweiler and Liberman, 1976) rescored the subjects' responses, this time ignoring order and giving credit to any correct item regardless of the order in which it was written down. This procedural change did not significantly alter the differences among the groups with regard to the factor of phonetic confusability. Finally, we have evidence from an altogether new study (Mark, Shankweiler, Liberman and Fowler, 1977) that good readers are more adversely affected by rhyming items than backward readers in a recognition memory experiment that entirely avoids the requirement of ordered recall.

All of our findings on language deficits in poor readers support Bakker's (1972) claim that in tests of perception and retention of serial order information, the verbal or nonverbal nature of the task requirements is crucial. In our view, there is ample reason to suppose that phonetic coding processes, and not merely length of memory span or temporal order perception, must be taken into account in order to find the causes of reading backwardness. From that perspective, poor serial recall is a symptom of difficulties in phonetic coding, not an independent deficit.

CONCLUSION

To summarize, our findings do not support the belief that a subclass of those specifically backward in reading, the dyslexics, can be differentiated from other poor readers on the basis of a high rate of reversal errors. Although some dyslexics showed orientational and directional biases that are absent in most poor readers, neither those children classified as dyslexics nor other poor readers typically displayed a high proportion of reversals as compared with other errors. Moreover, the difficulties manifested in the common error pattern are chiefly outside the domain of visual perception. They are language-related and are not specific to the visual perception of language. The difficulties of poor readers appear to reflect the inaccessibility of the phonetic segmentation of spoken language, inability to adopt an efficient coding strategy for operations involving short-term memory, and failure to grasp the complex nature of the spelling system of English. Since the difficulties of learning to read interact with the structural peculiarities of particular languages and the way those structures are manifested in the writing system, we must suppose that important work remains to be done in crosslanguage comparisons of children's reading errors. How these linguistic factors may influence the reading behavior of dyslexics is likely to be a productive question for future investigation.

REFERENCES

Baddeley, A. D. (1968) How does acoustic similarity influence short-term memory? Quarterly Journal of Experimental Psychology 20, 249-264.

Bakker, D. J. (1972) Temporal Order in Disturbed Reading. (Rotterdam: Rotterdam University Press).

Benton, A. L. (1975) Developmental dyslexia: Neurological aspects. Advances in Neurology, vol. 7, ed. by W. J. Friedlander. (New York: Raven Press).

Conrad, R. (1964) Acoustic confusions in immediate memory. British Journal of Psychology 55, 75-84.

Conrad, R. (1972) Speech and reading. Language by Ear and by Eye: The Relationships Between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).

Fischer, F. W., I. Y. Liberman and D. Shankweiler. (in press) Reading reversals and developmental dyslexia. Cortex.

Fowler, C. A., I. Y. Liberman and D. Shankweiler. (1977) On interpreting the error pattern in beginning reading. Language and Speech 20, 162-173.

Liberman, I. Y., D. Shankweiler, F. W. Fischer and B. Carter. (1974) Explicit syllable and phoneme segmentation in the young child. Journal of Experimental Child Psychology 18, 201-212.

Liberman, I. Y., D. Shankweiler, A. M. Liberman, C. A. Fowler and F. W. Fischer. (1977) Phonetic segmentation and recoding in the beginning reader. In Toward a Psychology of Reading: The Proceedings of the CUNY Conferences, ed. by A. S. Reber and D. L. Scarborough. (Hillsdale, New Jersey: Lawrence Erlbaum Associates).

Liberman, I. Y., D. Shankweiler, C. Orlando, K. S. Harris and F. Bell-Berti. (1971) Letter confusions and reversals of sequence in beginning reading: Implications for Orton's theory of developmental dyslexia. Cortex 7, 127-142.

Mark, L. S., D. Shankweiler, I. Y. Liberman and C. A. Fowler. (1977) Phonetic

recoding and reading difficulty in beginning readers. Memory and Cognition 5, 623-629.

Monroe, M. (1932) Children Who Cannot Read. (Chicago: University of Chicago Press).

Orton, S. T. (1937) Reading, Writing and Speech Problems in Children. (New York: Norton).

Shankweiler, D. and I. Y. Liberman. (1972) Misreading: A search for causes. In Language by Ear and by Eye: The Relationships between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).

Shankweiler, D. and I. Y. Liberman. (1976) Exploring the relations between reading and speech. In Neuropsychology of Learning Disorders: Theoretical Approaches, ed. by R. M. Knights and D. J. Bakker. (Baltimore: University Park Press).

Vellutino, F. (1977) Alternative conceptualizations of dyslexia: Evidence in support of a verbal-deficit hypothesis. Harvard Educational Review 47 (3), 334-354.

Weber, R. (1970) A linguistic analysis of first-grade errors: A survey of the literature. Reading Research Quarterly 4, 96-119.

Articulatory Units: Segments or Syllables?*

Thomas Gay+

ABSTRACT

This paper reviews present models of phoneme- and syllable-based models of articulatory programming and the physiological studies of speech production relevant to them. It then goes on to describe the results of some more recent research that is used to refine earlier formulations.

INTRODUCTION

The question of whether the motor input to the speech string is organized in terms of phoneme-size or syllable-size units has been an often studied, yet unresolved, issue in physiological speech research for a number of years. From an experimental point of view, the major obstacle to a solution is the nature of the speech signal that is available for observation. The speech string that presumably enters the articulatory mechanism as a set of discrete phonological units, emerges at the phonetic level as a continuously varying, highly encoded stream. The effect of this encoding can be observed in the production of a given phone in the form of a temporal spreading of its features to, or coarticulation with, adjacent phones. More importantly, these coarticulation effects are quite extensive, spreading beyond both segmental and syllabic boundaries, further obscuring the identity of the basic unit. Nonetheless, formal models of speech programming exist and a body of physiological data is relevant to them. This paper will review these models and a number of experimental findings related to them. It will then go on to describe the results of more recent research that will be used to refine the earlier formulations.

---

*A version of this paper was presented at the Symposium on Segment Organization and the Syllable, Boulder, Colorado, 21-23 October, 1977. It will appear in Syllables and Segments, ed. by Alan Bell and Joan B. Hooper, Amsterdam: North Holland Publishing Company, 1978.

+Also University of Connecticut Health Center, Farmington.

Acknowledgment: The comments and suggestions of, and assistance in collecting additional EMG data from Professor Björn Lindblom and Dr. James Lubker, Department of Phonetics, Institute of Linguistics, Stockholm University, are gratefully acknowledged. This research was supported by grants from the National Institute of Neurological and Communicative Disorders and Stroke (NS-10424), and the National Science Foundation (BNS-7616954).

[HASKINS LABORATORIES: Status Report on Speech Research SR-54 (1978)]

Existing Models

Coarticulation is usually defined as allophonic variation of a phone due to changes in its phonetic environment. These variations, which spread bidirectionally from right-to-left and left-to-right, arise from two different sources. Anticipatory (right-to-left) coarticulation effects are essentially timing effects: for a given segment, movements toward some parts of a feature target begin before others. Carryover (left-to-right) coarticulation effects, on the other hand, are usually considered to be mechano-inertia effects, and exist in the form of variability in target (or target feature) positions as a function of different preceding contexts. Both anticipatory and carryover coarticulation effects appear regularly in speech, and an explanation of both is needed for a general theory of speech production. However, because anticipatory effects originate at an early planning stage of the process, while carryover effects appear after the fact, so to speak, an explanation of anticipatory coarticulation is more central to the question of programming units.

In considering the nature of the articulatory programming unit underlying complex syllable constructions in Russian, Kozhevnikov and Chistovich (1965) proposed the general concept of the articulatory syllable. Using an electro-resistance measuring technique, Kozhevnikov and Chistovich tested their hypothesis by studying the onset of lip rounding (as reflected by lip protrusion) for the vowel /u/ placed in a number of CV, CCV, and CCCV contexts. Each of the segments in the consonant string was unmarked for labiality. Their data showed that the onset of rounding for the vowel usually began during the production of the first consonant in the string even if a syllabic boundary appeared within the string. Since the position of the syllable boundary was irrelevant to the timing of the protruding gesture, Kozhevnikov and Chistovich advanced the notion that the basic articulatory programming unit in speech was of a CV form with C corresponding to any number of consonants (C_1 - C_n), and V corresponding to a vowel. Kozhevnikov and Chistovich further suggested that the motor instructions for each segment within the syllable would be sent out simultaneously, unless the syllable contained competing articulations, in which case the instructions would be issued in sequence.

A second formal model of articulatory programming was proposed by Henke (1966). Using a computer simulation program based on midsagittal x-ray films of the vocal tract, Henke developed a dynamic model that describes the state of the vocal tract at successive points in time. The input to the model is a discrete phonemic string. Anticipatory features associated with a down-stream segment are added to the state of the model at any moment in time by means of a "look ahead" mechanism. This mechanism scans future segmental inputs and issues commands for the immediate attainment of the feature targets of those segments that would not interfere with the attainment of intervening articulations.

Thus anticipatory coarticulation of lip rounding in a CCCV sequence can be explained by the Kozhevnikov and Chistovich model and by the Henke model as well. In Kozhevnikov and Chistovich's model, the onset of lip rounding is considered as part of the set of simultaneous motor commands issued for the production of the entire articulatory syllable. In the Henke model, since the three consonants are unmarked for rounding, the "look ahead" mechanism

searches for the next segment marked for rounding, and issues commands for the rounding gesture to begin during the unspecified consonant string.

While both the articulatory syllable and phoneme-by-phoneme models can explain most of the anticipatory coarticulation effects observed for lip rounding, most investigators prefer to interpret their results in terms of the Henke (1966) model primarily because of its simplicity and greater explanatory powers. While Henke's model specifies a simple phoneme-by-phoneme input, Kozhevnikov and Chistovich's is built around an unnatural and counterintuitive syllable that bears no simple correspondence to common linguistic or phonetic units.

The Forward Spreading of Anticipatory Coarticulation

Anticipatory coarticulation has been studied primarily in terms of three different articulatory features: lip rounding for a rounded vowel, tongue body movements for a postconsonantal vowel in a VCV sequence, and velar lowering for a nasal consonant, with most of the research questioning how far in advance of the particular segment these anticipatory movements begin.

In extending the observations of Kozhevnikov and Chistovich (1965) to American English, Daniloff and Moll (1968) showed that the onset of lip rounding for the vowel /u/ can begin across as many as four consonant segments ahead of the vowel. In their cinefluorographic experiment, the onset of lip protrusion for /u/ was studied for a number of mono- and disyllabic single and two word utterances embedded in sentence frames. *Onset of lip rounding* usually began with the first consonant in the string, and was not affected by the position of either syllable or word boundaries that appeared in the string. Similar anticipatory lip rounding effects have been demonstrated by Lubker, McAllister and Carlson (1975) at the EMG level, and Benguerel and Cowan (1974) at the movement level. Lubker, McAllister and Carlson's data showed that the onset of EMG activity for the orbicularis oris muscle (a primary lip rounding muscle) associated with /u/ began with the first consonant in the string, and as early as 600 msec prior to the onset of the vowel. In the Benguerel and Cowan study, the onset of lip protrusion for /u/ in a number of similar utterances for speakers of French likewise usually appeared at the time of the first consonant. Benguerel and Cowan also observed that the lip rounding gesture sometimes began as early as the preceding vowel segment. This finding was used to argue specifically against the Kozhevnikov and Chistovich (1965) model; the observed VC coarticulation was inconsistent with the concept of an open CV syllable. However, Benguerel and Cowan did not specify the actual point in time during the vowel when these movements occurred, leaving open the possibility that the movement began during the vowel-consonant transition portion of the vowel.

Another weakness of the Kozhevnikov and Chistovich model is that it does not predict coarticulation effects for certain consonant features--velar lowering, for example--and is not general enough to explain articulatory coarticulation in simple V or VC sequences, strings where anticipatory coarticulation has been shown to exist. For example, in a spectrographic study of coarticulation in Swedish VCV sequences, Öhman (1965) suggested that the variability observed in transition movements from the first vowel to the intervocalic consonant could be predicted by the formant frequencies of the

second vowel. This led Öhman to conclude that the consonant gesture in a VCV sequence is simply superimposed on a basic vowel-to-vowel substrate. In other words, anticipatory movements toward the second vowel can begin independently of those toward the consonant. Examples of anticipatory coarticulation of velar lowering have also been reported in the literature. Moll and Daniloff (1971) showed that, in a CVVN sequence, velopharyngeal opening for the final nasal usually began during the production of the first vowel in the sequence, that is, two segments in advance of the nasal consonant. McClean (1973) observed similar patterns of velar lowering. In his data, anticipatory velar opening for a final nasal in a CVVN sequence usually began with the first vowel in the sequence, unless the two vowels were separated by a marked junctural boundary.

The studies reviewed above, among others, suggest that articulatory encoding is a complex phenomenon whose effects can spread across several adjacent segments, ungoverned by simple linguistic or phonetic rules. Most support, either explicitly or implicitly, Henke's (1966) articulatory model on the basis that the look-ahead mechanism central to this model can explain the observed coarticulatory variations. However, in recent studies, both cinefluorographic and electromyographic evidence was used to argue against the pervasiveness of these effects in general and the operation of Henke's model in particular (Gay, 1975; Gay, 1977a, 1977b).

The Forward Limits of Anticipatory Coarticulation

The two experiments described below were undertaken for the purpose of studying, in greater detail than had been done before, the coordination of articulatory gestures in VCV sequences, at both the articulatory movement and EMG levels.

In the cinefluorographic experiment (Gay, 1977a), conventional high speed (60 fps) lateral view x-ray films were obtained from two subjects who produced various VCV utterances that contained the vowels /i,a,u/ and the consonants /p,t,k/ in all possible combinations. Articulatory movements were tracked by recording the positions, frame-by-frame, of 2.5 mm diameter lead pellets that had been attached to the upper and lower lips, jaw, and several locations along the surface of the tongue, relative to a reference pellet attached at the embrasure of the upper central incisors. This experiment was designed for the specific purpose of exploring the question of whether, in a VCV sequence, an intervening consonant constrains the movements of the articulators, in particular the tongue body and lips, from one vowel to the other; in other words, is the movement from one vowel to another in the form of a simple substrate (Öhman's hypothesis) or is it somehow locked to the consonant, and is the lip rounding gesture for the postvocalic rounded vowel likewise constrained by the intervocalic consonant?

The dynamic properties of articulatory movements in a VCV sequence are illustrated in Figure 1 for an utterance where the intervocalic consonant is /p/. This figure shows the movement tracks of the tongue body, lips and jaw in the height dimension for the sequence /kipap/ as produced by two different speakers. Each track is graphed from discrete points measured every film frame, that is, at approximately 17-msec intervals. Measurements begin during the closure period of the initial /k/ and end at the time of closure for the

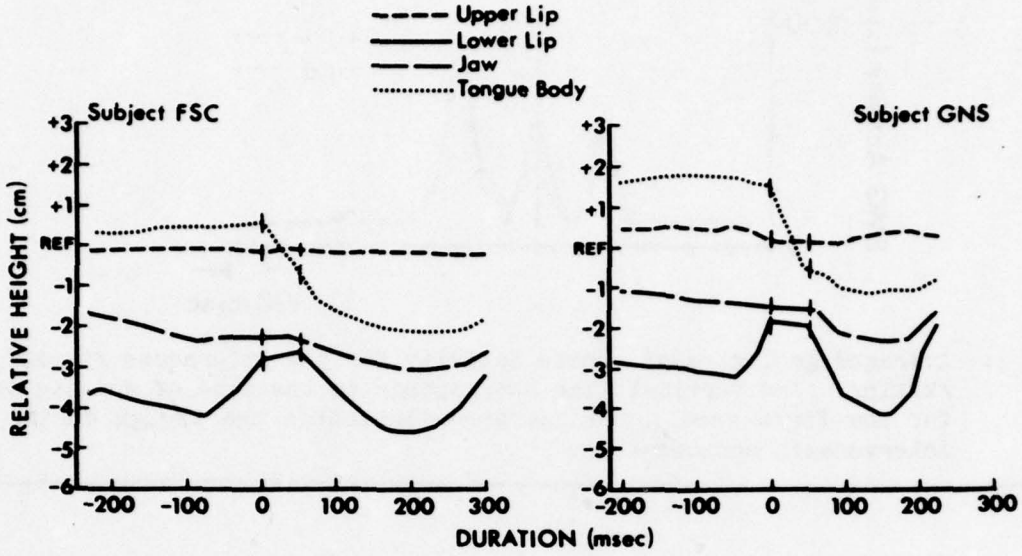


Figure 1: Movement tracks (height) for the utterance /kipap/, two speakers. The vertical lines indicate the times of lip closure and release for the intervocalic /p/.

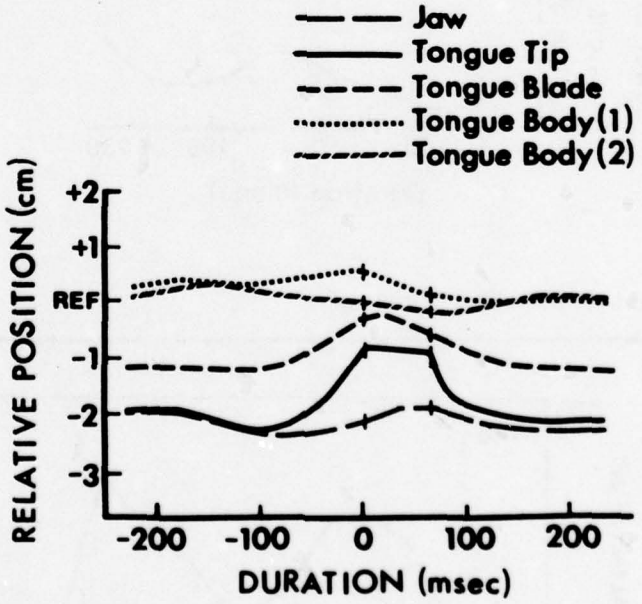


Figure 2: Movement tracks for the utterance /kitip/.

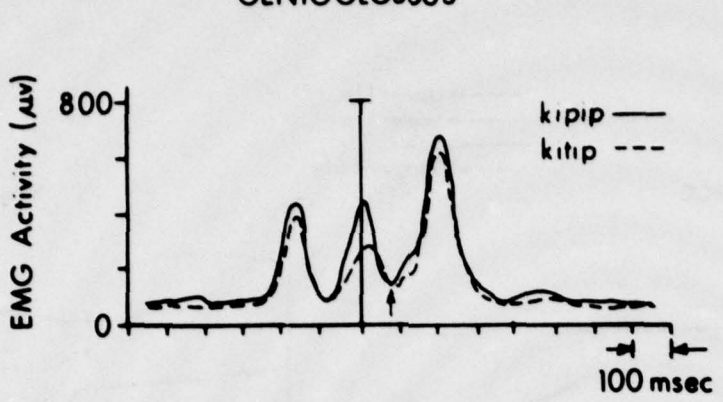


Figure 3: Averaged genioglossus muscle activity for the utterances /kipip/ and /kitip/. The vertical line corresponds to the time of voicing onset for the first vowel, and the arrow indicates the trough during the intervocalic consonant.

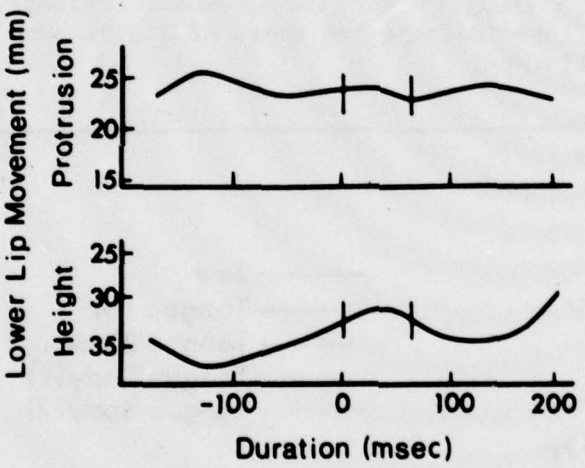


Figure 4: Movement tracks for /kutup/.

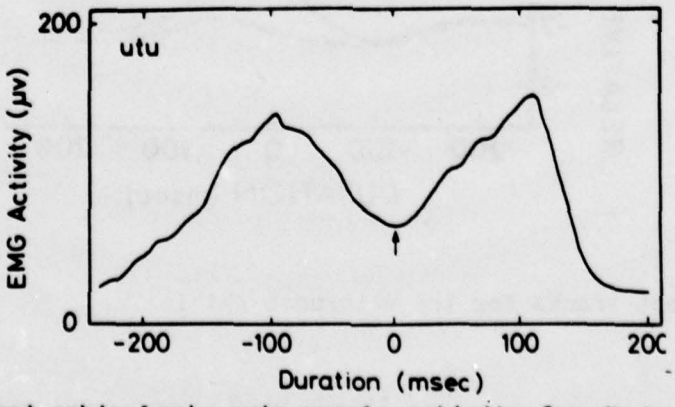


Figure 5: Averaged orbicularis oris muscle activity for /kutup/.

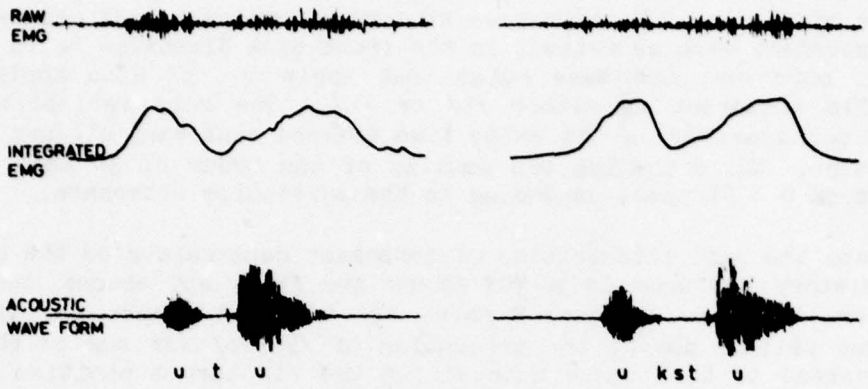


Figure 6: Orbicularis oris muscle activity for /utu/ and /ukstu/.

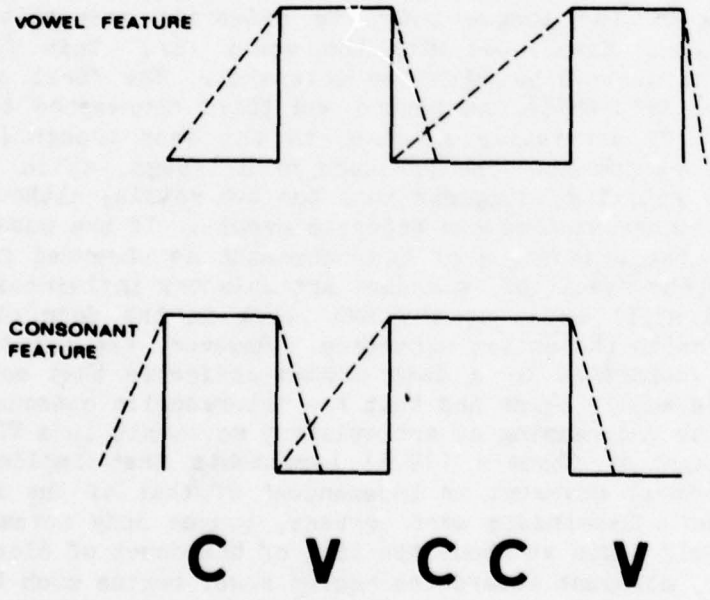


Figure 7: Schematic representation of coarticulatory field for vowel and consonant features.

final /p/. The 0 on the abcissa corresponds to the time of closure for the intervocalic vowel. This figure illustrates the general finding that an intervocalic consonant affects the timing of the movements of the tongue body from vowel to vowel. The movement of the tongue body from the first vowel to the second vowel does not begin until after closure for the intervocalic consonant is completed. This was found to be a salient feature in the production of all the VCV utterances studied. Consonant constraints on vowel-to-vowel movement were as evident in the front-back dimension as in the height dimension; moreover, the same rules that apply to /p/ also apply when the intervocalic consonant is either /t/ or /k/. The only variability in the timing effect appeared in the delay time between consonant closure and tongue body movement. While the lag was usually of the order of 30 msec, it varied anywhere from 0 - 60 msec, depending on the particular utterance.

Perhaps the best illustration of consonant constraints on the programming of articulatory gestures is a VCV where the first and second vowels of the sequence are the same. Figure 2 shows the movement tracks for the jaw and four tongue pellets during the production of /kitip/ for one of the subjects (FSC). Instead of the tongue maintaining the /i/ target position during the consonant, the tongue blade and both tongue pellets show continuous movement during the entire consonant gesture. The blade and anterior tongue body pellet (Pellet 1) appear to shadow movements of the tip, while the posterior tongue body pellet (Pellet 2) moves in the opposite direction, either in a facilitory gesture or towards a tongue body consonant target.

The effects of the intervocalic consonant on the organization of the vowel gesture in a symmetrical VCV are even more evident at the EMG level (Gay, 1975). The average EMG activity of the genioglossus muscle for the sequences /kipip/ and /kitip/ as produced by one of the subjects (FSC) of the x-ray experiment, is illustrated in Figure 3. The genioglossus muscle, which comprises the bulk of the tongue body, is primarily responsible for the protruding and bunching associated with the vowel /i/. This figure shows three separate peaks associated with the utterance. The first peak corresponds to the initial /k/, while the second and third correspond to the first and second vowels. Of particular interest is the deep trough (arrow) that separates the two vowel peaks. The presence of a trough, which signifies a cessation of muscle activity, suggests that the two vowels, although phonetically identical, are organized as two separate events. If the movement of the tongue body during the production of the consonant as observed in the x-ray data (Figure 2) was the result of secondary articulatory influences, positional constancy would still exist at the EMG level in the form of one broad genioglossus peak across the entire utterance. However, the existence of two distinct EMG peaks separated by a deep trough indicates that each vowel is marked by a separate muscle pulse and that the intervocalic consonant plays an inportant part in the programming of articulatory movements in a VCV sequence. These data argue against Öhman's (1965) hypothesis that implies that the timing of vowel-to-vowel movement is independent of that of the intervocalic consonant. If Öhman's hypothesis were correct, tongue body movements toward the second vowel would begin at about the time of the onset of closing for the consonant. However, movement toward the second vowel begins much later, up to 60 msec after closure for the consonant has already been completed. In other words, the onset of anticipatory tongue body movements for the second vowel in a VCV sequence seems to be limited by the onset of closure of the preceding

consonant.

In addition to placing constraints on the movements of the tongue body from one vowel to another in a VCV, an intervocalic consonant also affects the onset of lip rounding for a rounded second vowel. These constraints are observable at both the EMG and articulatory movement levels (Gay, 1975; Gay, 1977b). In those cases where a rounded vowel appears in a postconsonantal position, the rounding gesture (as reflected by lip protrusion), like tongue body movement, does not begin until after closure for the intervocalic consonant is completed. This is true even for the most sensitive case, namely, a symmetrical VCV containing the same rounded vowels. Figure 4 shows the movement tracks of lower lip height and lower lip protrusion plotted against the same baseline for the sequence /kutup/. Even in this example, it is evident that the rounding feature of the first vowel is not continuous through the consonant. Rather, what appears to be an additional, although small, closing and protruding gesture is superimposed on the rounding pattern. This can be seen as a perturbation in both the lip height and lip protrusion curves during the time of consonant production.

The discontinuity of lip rounding during consonant production is more obvious at the EMG level. Figure 5 shows the corresponding EMG data for the orbicularis oris muscle during the production of the same utterance. The envelope represents the average of some 16 tokens of the utterance. The 0 on the time scale corresponds to the time of voicing offset of the first vowel. This figure shows a deep trough in the EMG envelope during the time of consonant production. Again, the presence of a trough signifies an interruption of muscle activity corresponding to the time of consonant production. While these findings are similar to most of the experimental findings that showed the onset of the rounding gesture to occur during the production of a preceding consonant, or consonants (Daniloff and Moll, 1968; Lubker, McAllister and Carlson, 1975, among others), the presence of the trough argues against the interpretation that the onset of rounding is controlled by a look-ahead mechanism of the type proposed by Henke (1966). Since the first vowel is marked for rounding while the intervocalic /t/ is unspecified, Henke's model would predict that the rounding feature would be retained during the production of the consonant. At the EMG level, this would be reflected by a single broad envelope from the beginning of the first /u/, through the consonant, to the end of the second /u/. Obviously, however, this does not happen. The two vowels are each marked by a separate and distinct muscle pulse, with the onset of the lip rounding gesture for the second vowel constrained, like movements of the tongue body, by the time of closure of the intervocalic consonant.

Because the onset of lip rounding in a VCV sequence seems to occur considerably closer to the vowel than it does in a CCV sequence, the question arises whether the two types of sequences are governed by different rules. To answer this question, additional EMG recordings were recently obtained from the orbicularis oris muscle during the production of a number of VCV, VCCV, and VCCCV sequences containing the rounded vowel /u/ in both pre- and post-consonantal positions. An illustration of the findings for two extreme cases /utu/ and /ukstu/, as produced by an English-speaking subject, appear in Figure 6. This figure shows both the raw and integrated EMG signals for the upper lip electrode plotted against the acoustic waveform. The integrated EMG

envelope for /utu/ replicates the earlier finding: a trough separates the two vowel peaks. More interesting, however, is the finding that the patterns for the VCCV and VCCCV sequences, illustrated in this figure by /ukstu/, are also characterized by a trough. The first vowel is marked by a burst of muscle activity that ceases when the consonant appears, but resumes almost immediately once the consonant is underway. The pattern illustrated here typifies the slow increase in rounding activity that is usually associated with the second vowel. These findings, which have also been observed for Swedish-speaking subjects by Lubker[1] provide a convincing illustration that the field of anticipatory lip rounding for the second vowel in any $VC_1 - C_nV$ sequence seems to be limited by the forward boundary of the consonant category preceding the vowel.

Thus, both electromyographic and cinefluorographic evidence suggest that the relative timing of articulatory movements toward a vowel in $VC_1 - C_nV$ sequences is affected by the prevocalic consonant(s), even if the consonantal features are not contradictory to those movements. The consonant affects both the tongue body movements toward, and lip rounding gesture for the postconsonantal vowel, in that anticipatory movements toward the second vowel do not begin earlier than the time of closure for the consonant. These findings argue against the appearance of anticipatory movements across phonetic categories and the operation of a look-ahead mechanism for the control of these features.

Do these findings suggest that we turn from a phoneme-based articulatory model to an articulatory syllable model? Not necessarily; the reasons for rejecting the concept of a CV syllable as the basic unit in articulatory programming are convincing ones. Rather, these findings suggest that if the string is organized on a phoneme-by-phoneme basis, the input must be more tightly controlled than Henke's (1966) model specifies. Since the forward extent of anticipatory coarticulation seems to be limited by the forward boundary of the preceding phonetic category, it is reasonable to suggest, that as a general rule, the size of the anticipatory field can be defined solely in terms of the size of that phonetic category.

An illustration of how this rule would operate in speech is shown schematically in Figure 7. The blocks represent successive vowel and consonant features (solid lines) of a hypothetical CVCCV sequence. The coarticulatory field for each segment (dashed) line extends bidirectionally (anticipatory effects to the left and carryover effects to the right) across the segment boundaries. For vowel features such as lip rounding, anticipatory movements can begin well ahead of the vowel and across both syllable and word boundaries, but not across phonetic category boundaries. These field effects are essentially the same for consonants: anticipatory movements associated with consonant production are limited to the field of the preceding vowel category. The duration of the anticipatory field is related solely to the duration of the phonetic category. If the preceding phonetic category contains several segments, the size of the anticipatory field will be larger and the onset of the anticipatory movements will be earlier than if the

---

[1]Lubker, J: personal communication.

preceding phonetic category contains only a single segment.

As expressed above, this formulation is similar to Henke's (1966) model, divested of its temporally unspecified look-ahead mechanism. However, it is based largely on data relevant to a single, somewhat peripheral articulatory feature, namely, lip rounding. Obviously, additional data are needed before these findings can be generalized. In particular, rules for the timing of tongue body and jaw movements in consonant clusters, and the effects of stress and speaking rate on the temporal properties of these segmental gestures must be established before the picture can be completed. However, at this point, it is reasonable to speculate that the motor input to the speech mechanism seems to operate by simple rules on phoneme-sized units and within a specifiable temporal field.

REFERENCES

Bengurel, A.-P. and H. Cowan. (1974) Coarticulation of upper lip protrusion in French. Phonetica 30, 41-55.

Daniloff, R. and K. Moll. (1968) Coarticulation of lip rounding. Journal of Speech and Hearing Research 11, 707-721.

Gay, T. (1975) Some electromyographic measures of coarticulation in VCV utterances. Proceedings of Firth Phonetics Symposium, Essex, 15-29.

Gay, T. (1977a) Cinefluorographic and electromyographic studies of articulatory organization. In Dynamic Aspects of Speech Production, ed. by M. Sawashima and F. S. Cooper. (Tokyo: University of Tokyo Press), 85-102.

Gay, T. (1977b) Articulatory movements in VCV sequences. Journal of the Acoustical Society of America 62, 183-193.

Henke, W. (1966) Dynamic articulatory model of speech production using computer simulation. Doctoral dissertation, M.I.T.

Kozhevnikov, V. and L. Chistovich. (1965) Rech', Artikulyatsiya, i Vospriyatiye. Translated as Speech: Articulation and Perception. (Washington, D.C.: Joint Publications Research Service), 30 pp., 543.

Lubker, J., R. McAllister and J. Carlson. (1975) Labial co-articulation in Swedish: a preliminary report. In Procedings of Speech Communication Seminar, ed. by G. Fant. (Stockholm: Almqvist and Wiksell), 55-64.

McClean, M. (1973) Forward coarticulation of velar movement at marked junctural boundaries. Journal of Speech and Hearing Research 16, 286-296.

Moll, K. and R. Daniloff. (1971) Investigation of the timing of velar movements during speech. Journal of the Acoustical Society of America 50, 678-684.

Öhman, S. (1965) Coarticulation in VCV utterances: Spectrographic measurements. Journal of the Acoustical Society of America 39, 151-168.

Selective Anchoring and Adaptation of Phonetic and Nonphonetic Continua

Helen J. Simon[+] and Michael Studdert-Kennedy[++]

ABSTRACT

A series of four experiments compared the effects of unequal probability anchoring and selective adaptation on phonetic and nonphonetic judgments. The basic stimulus series was a synthetic stop consonant continuum ranging from /b/ to /d/. On this continuum were superimposed covariations in fundamental frequency, intensity or vowel. In each experiment subjects listened to identical test tapes under two judgment conditions: place of articulation and pitch or loudness or vowel judgments. The two types of judgment were significantly dissociated under both anchoring and adaptation paradigms, thus demonstrating that the former may be no less selective than the latter. From this and other evidence, it was concluded that the two paradigms are, in principle, equivalent, and that the main factors in speech adaptation effects are peripheral fatigue and central auditory contrast. If the selective processes of fatigue and contrast are taken to reflect functional channels of analysis rather than the operation of feature detectors, the same broad processes can be seen at work in both *speech and nonspeech* adaptation.

INTRODUCTION

Over the past five years, several dozen papers have reported studies of the selective adaptation of speech sounds. The series began with a paper by Eimas and Corbit (1973). They asked listeners to categorize members of a synthetic voice onset time (VOT) continuum (Lisker and Abramson, 1964) and demonstrated that the perceptual boundary between voiced and voiceless catego-

---

[+]Arizona State University, Tempe, Arizona.

[++]Also at Queens College and the Graduate Center, City University of New York.

Acknowledgment: The experiments reported here are drawn from a thesis by the first author submitted to the Graduate Center, City University of New York in partial fulfillment of the requirements for the Ph.D. degree. The work was supported in part by NICHD Grant HD-01994 to Haskins Laboratories. Portions of the paper were completed while the second author was on sabbatical leave as a research fellow at the Center for Interdisciplinary Research, University of Bielefeld, Germany. For advice, discussion and comments we thank Peter Bailey, Joanne Miller, David Pisoni, Bruno Repp and Quentin Summerfield.

[HASKINS LABORATORIES: Status Report on Speech Research SR-54 (1978)]

ries along that continuum was shifted by repeated exposure to (that is, adaptation with) either of the endpoint stimuli: there was a decrease in the frequency with which stimuli close to the original boundary were assigned to the adapted category and a consequent shift of the boundary toward the adapting stimulus. Since the effect could be obtained on a labial VOT continuum after adaptation with a syllable drawn from an alveolar VOT continuum, and vice versa, adaptation was clearly neither of the syllable as a whole nor of the unanalyzed phoneme, but of a feature within the phonemic segment. Eimas and Corbit therefore termed the adaptation "selective" and attributed their results to the fatigue of specialized linguistic feature detectors and the relative "sensitization" of opponent detectors. Subsequent studies have replicated the results for VOT and have extended them to other phonetic feature oppositions, such as place and manner of articulation. For reviews, see Ades (1976), Cooper (1975) and Eimas and Miller (in press).

These later studies have generally continued to favor a detector fatigue model, but have tended to modify the hypothesized level of adaptation by attributing the effects to acoustic rather than linguistic feature detectors. Thus, in a recent version of the model, Eimas and Miller (in press) propose "...multiply-tuned feature detectors, each...sensitive to a range of complex acoustic information...sufficient to signal a single phonetic feature value...," and each having its "...greatest sensitivity...at some set of acoustic values that correspond to the modal acoustic consequences of articulating that feature value in a particular set of conditions...."

Attractive as such an account may be for those concerned with the biology of language in general, and with the ontogeny of speech perception in particular, there are both broad and narrow grounds for scepticism. Broadly, the hypothesis that selectively tuned feature detectors segment the speech signal at an early stage of its processing into sets of invariant properties, is not easy to mesh with the accumulated evidence that cues to phonetic structure are highly variable and are often distributed dynamically across an entire syllable (for example, Cooper, Delattre, Liberman, Borst and Gerstman, 1952; Liberman, Cooper, Shankweiler and Studdert-Kennedy, 1967; Shankweiler, Strange and Verbrugge, 1976; Studdert-Kennedy, 1976a; Dorman, Studdert-Kennedy and Raphael, 1977; and for recent discussions bearing directly on the question of feature detectors, see Repp, 1977; and Repp, Liberman, Eccardt and Pesetsky, 1978). Moreover, evidence to support the several assumptions of the model comes entirely from speech adaptation studies. While there is no question that these studies have persuasively demonstrated channels of feature analysis in speech perception, they scarcely warrant literal interpretation of a physiological metaphor. For although complementary (or opponent) detectors may be plausibly invoked to account for the perception of, say, color or movement, the function of such detectors in the perception of, say, laryngeal periodicity or formant transition is less easily imagined.

Among the narrower grounds for doubt are proliferating reports of contingent effects in adaptation studies (Ades, 1976): adaptation of initial stop consonants on both voicing (Cooper, 1974) and place of articulation (Bailey, 1973; Pisoni, Sawusch and Adams, 1975; Miller and Eimas, 1976) is now known to be contingent upon the following vowel, while adaptation of place of articulation has been found to depend also upon syllable position (Ades, 1974; Miller and Eimas, 1976) and fundamental frequency (Ades, 1977). As Eimas and

Miller (in press) explicitly recognize, the theoretical utility of selectively tuned feature detectors goes down as the number of contexts to which they must be tuned goes up.

A second reason for doubt is that the detector fatigue supposedly induced by adaptation has usually been inferred from shifts in response frequency rather than measured directly by threshold determination in the manner of, for example, Kay and Matthews (1972) who traced the tuning curves of human auditory channels sensitive to specific ranges of frequency modulation. Recent work by Miller (1977), Miller, Eimas and Root (1977) and Sawusch (1976, 1977) has begun to fill this gap, and we will take account of their results in the final discussion, but the only other attempt to estimate sensitivity changes directly, in a speech adaptation study, produced results precisely the opposite of those predicted by a fatigue model. Cooper, Ebert and Cole (1976) used a magnitude estimation procedure (Durlach and Braida, 1969) to assess d' values, before and after adaptation, for pairs of stimuli along a labial stop-to-glide continuum. Although they found a general decrease in sensitivity after adaptation, by far the largest decrease was at the phoneme boundary rather than at the adapted stimulus, as a fatigue model would predict. In fact, for the two pairs of stimuli that included the endpoint adaptors, Cooper et al. (1976) reported an increase in sensitivity, which, though not tested for significance, led them to countenance the "...strong possibility...that the increased sensitivity following adaptation for a pair of stimuli including or neighboring the adapting stimulus reflects the listener's heightened ability to utilize the adapting stimulus as an anchor, or reference stimulus..." (Cooper et al., 1976, p. 103).

Such a possibility is exactly what the present experiments were designed to explore. Our starting point was a study by Sawusch and Pisoni (1973). They collected identification functions for a series of synthetic speech sounds varying in VOT from /b/ to /p/ and for a series of tones varying in intensity from 60 to 80 dB SPL ("soft" to "loud"). In one condition (control) all stimuli occurred equally often. In the other (anchor) a particular endpoint stimulus occurred twice as often as other stimuli in the series (2:1 ratio)[1]. The result was that in the anchor condition the category boundary for the tones shifted toward the anchoring stimulus, while the phonetic boundary remained stable. In a second experiment, Sawusch, Pisoni and Cutting (1974) used the same anchoring procedure (at a 4:1 ratio) on a synthetic continuum in which the stimuli varied simultaneously in both place of articulation (/b/ to /d/) and fundamental frequency (F_0) ("high" to "low"). Once again the nonphonetic boundary shifted, while the phonetic boundary remained stable. Since the authors believed that the effect of anchoring is to bias a subject's responses rather than his percepts, they inferred that stop consonants are largely immune to response bias. They attributed this immunity to the listener's having an internal standard for the stop consonant

---

[1]The term "anchor" has several usages (see Parducci, 1974, p. 132). Most commonly it is used for a reference or standard stimulus outside the series being judged. However, we here follow Sawusch and Pisoni (1973) and use the term for a stimulus within the series to be judged, but assigned a higher frequency of occurrence.

and concluded that the effects of selective adaptation to which stop consonants are not immune were therefore sensory, or perceptual.

However, this argument reaches the right conclusion for the wrong reason. Not all anchoring effects reflect response bias (see, for example, Helson and Kozaki, 1968; Helson, 1971), and it seems unlikely that two procedures, differing as little as do anchoring and adaptation, would engage qualitatively different mechanisms in judgments of the same continuum. The main difference between the procedures is simply in the number and distribution of anchor or adaptor repetitions. In adaptation, the adaptor typically occurs many times at short interstimulus intervals (ISI) in a single block before the test stimuli. In anchoring, the anchor occurs less often and with a longer ISI, its repetitions being scattered randomly among the test stimuli. The two procedures lie at opposite ends of a continuum of anchor/adaptor energy concentration, so that the differences between the consonantal results in the two types of experiment may well be of degree rather than of kind.

The following experiments therefore address three questions: (1) Is there a reliable dissociation in the effects of anchoring between judgments of phonetic and judgments of nonphonetic continua? (2) Is there a similar dissociation in the effects of adaptation? (3) Can a single, unified account be developed for the effects of both experimental paradigms on both types of continua? Experiments I, II and III approach the first question by comparing the effects of anchoring on the identification of a synthetic stop consonant continuum with those on the identification of fundamental frequency, intensity and synthetic vowel continua. The two types of stimulus variation in each experiment covary, that is to say, are perfectly correlated and are carried simultaneously on the same series of syllables. Experiment IV approaches the second question by comparing the effects of both anchoring and adaptation on the identification of covarying synthetic stop consonant and fundamental frequency continua. We reserve the third question for our concluding discussion.

EXPERIMENT I

The first experiment replicated the study of Sawusch et al. (1974) with one difference. These authors used a seven-step synthetic stop place of production continuum (/b/ to /d/) and paired each stimulus value along the continuum with each of seven variations in F_0. They thus produced forty-nine stimuli in which formant structure and F_0 varied independently. In the present study, each stimulus along a seven-step synthetic stop place of production continuum was assigned a different F_0 contour. This yielded only seven stimuli in which formant structure and F_0 were perfectly correlated. If under these conditions a shift in the fundamental frequency boundary was observed without a shift in the phoneme boundary, an even stronger case might be made for the dissociation of phonetic and nonphonetic judgments in an anchoring paradigm.

Method

Stimuli. A series of three-formant consonant vowel (CV) syllables was generated on the Haskins Laboratories parallel resonance synthesizer. The

series consisted of seven syllables, each with a duration of 300 msec. The final 260 msec was a steady-state portion appropriate to the American English vowel /æ/ and was identical in all syllables: 660 Hz for F_1, 1620 Hz for F_2 and 3026 Hz for F_3. The seven stimuli ranged perceptually from /bæ/ to /dæ/ in approximately equal steps in second- and third-formant transition starting frequencies (Table 1). For all seven stimuli, F_1 rose over the first 40 msec from a starting frequency of 234 Hz to its steady-state frequency of 660 Hz.

Fundamental frequency for the first 225 msec of each syllable varied from 114 Hz to 150 Hz in 6 Hz steps over the series from /b/ to /d/ and fell linearly from its initial value to 80 Hz during the last 75 msec of the syllable. Therefore, a particular fundamental frequency contour characterized each syllable along the phoneme continuum. Overall amplitude was attenuated linearly by 28 dB in the last 75 msec of each syllable.

The syllables were recorded on magnetic tape. They were then digitized, edited and stored on the Haskins pulse code modulation system (PCM) for subsequent tape preparation. Two identification tapes were prepared. On the first, the control tape, the seven syllables were recorded equally often: ten random permutations of the series to produce seventy stimuli. On the second, the anchoring tape, the anchor stimulus was recorded four times as often as the other six stimuli: ten random permutations of ten stimuli (four anchors plus six others) to yield a total of 100 stimuli. The anchor stimulus was the first syllable (/bæ/) described in Table 1, with an F_0 of 114 Hz. Since a differential effect of the two end points was not of interest in this experiment, only one end-point was used as an anchor.

The stimuli on both tapes were recorded on a Crown 800 tape recorder connected directly to the output of the PCM system, with a 2-sec interval between stimuli and a 10-sec interval after every tenth stimulus. A 1000 Hz calibration tone, set at the peak vowel amplitude on the VU meter of the tape recorder, was recorded at the beginning of each tape to permit uniform playback levels.

Procedure. The experiment called for identification of the two stimulus series, control and anchor, under two task conditions, pitch identification and phoneme identification. Each subject heard the same control (70 item) and the same anchor (100 item) tape twice. Only the instructions varied as the task changed. For the pitch task, subjects were told that they would hear a series of sounds presented on two tapes similar in every way save that the second had 100 stimuli, while the first had only seventy, and that while other aspects of the stimuli would vary, their task was simply to write "high" or "low" to identify the pitch. For the phoneme identification task, the subjects were given similar instructions, but were asked to identify each syllable as beginning with either /b/ or /d/. Each experimental tape was preceded by display and practice tapes.

The order of testing was the same for all subjects: pitch control, pitch anchor, phoneme control and phoneme anchor. The pitch task was deliberately given first, since the expected absence of an anchoring effect in the phoneme task would then have occurred in spite of the expected anchoring effect in the preceding pitch task.

All experimental tapes were reproduced binaurally from the output of an Ampex AG 500 tape recorder over calibrated Telephonics (TDH-39) matched headphones with a circumaural seal. The gain of the tape recorder playback for the 1000 Hz calibration tone was adjusted to give a voltage across the earphones equivalent to approximately 75 dB SPL re 0.0002 dyne/cm^2.

Subjects. Nine normal hearing (screened at 20 dB HTL re: ANSI 1969 standard) undergraduate students at Yale University participated in this experiment. All were native speakers of English, had no past history of speech or hearing problems, and were paid at a rate of $2.00 per hour. They were tested in a quiet room, either alone or in pairs, in a single session with a short break between the pitch and phoneme judgment conditions.

Results

Figure 1 displays the pitch and consonant identification functions for control and anchor conditions. Apart from the marked discontinuity in the pitch control function (discussed below), the most obvious features of the figure are the shallowness of the pitch functions--as compared with the steepness of the consonantal functions--and the clear leftward shift of the pitch anchor function--as compared with the trivial shift of the consonantal anchor function.

As a simple measure of the anchoring effect, we can compare the distribution of the two response categories over the stimulus continuum in the control and anchor conditions. If we assume an effect of contrast rather than assimilation, we can then predict a decrease in the total number of responses in the class to which the anchor is typically assigned. Since, in the anchoring condition, the anchor stimulus was presented more often than the other stimuli, and was almost always assigned to the same response class, responses to this stimulus were omitted from the totals. Table 2 therefore lists individual and mean numbers of "low" and "b" identification responses to the sixty presentations of the remaining stimuli under the two task conditions. For the pitch task there is a mean decrease of 7.2 "low" responses in the anchor condition; the decrease is significant by a one-tailed, matched pairs t-test ($t = 4.07$, $p < .005$). For the speech task there is an insignificant mean increase of 0.2 "b" responses in the anchoring condition.

A second, derived measure of an anchoring effect is the shift in the estimated boundary between response categories: here, a contrast effect would appear as a shift in the boundary toward the anchor stimulus. Normal ogives were fitted to the individual data by the method of least squares[2] (Woodworth,

---

[2]Although this procedure usually leads to the same conclusion as does a direct count of the decrease in the number of responses assigned to the anchor class, it also permits systematic measurement of response variability. Since the degree of response variability (that is, of stimulus ambiguity) may bear importantly on the interpretation of the results, and since boundary shifts are both easily read on the figures and have become standard measures in the adaptation literature, we report both direct and indirect measures throughout this paper.

1938), and the results are listed in Table 3. For the pitch task, the mean control and anchor boundaries (that is, means of the fitted ogives) are 3.63 and 3.24 respectively, a significant shift in the anchor condition of 0.34 continuum steps toward the anchor stimulus (matched pairs $t = 2.03$, $p < .05$, one-tailed). For the speech task, the mean control and anchor boundaries are 3.74 and 3.83 respectively, an insignificant shift in the anchor condition of 0.07 steps away from the anchor.

The individual and mean standard deviations (reciprocally related to the slopes) of the ogives fitted to the control data from the two task conditions are listed in Table 4. The speech standard deviations are remarkably consistent across individuals and every subject, except subject 8, gives a larger standard deviation on the pitch task than on the speech task. The mean pitch standard deviation of 1.16 is significantly larger than the mean speech standard deviation of 1.02 on a two-tailed matched pairs t-test ($t = 2.40$, $p < .05$). There is no significant Spearman coefficient of rank order correlation between control standard deviation and boundary shift for either pitch (rho = -.14) or speech (rho = .09).

The relatively large pitch standard deviations clearly result, in part, from the sharp discontinuity in the pitch control function. This discontinuity itself seems to have resulted from a simple contrast effect, precipitated by an accident of the randomized test order. By a chance not noticed until after the data had been gathered, stimulus 3 was always preceded by either stimulus 1 or stimulus 2, both of which were, in this position, always identified as "low." Stimulus 4, on the other hand, was always preceded by stimulus 5, 6 or 7 which on 96 percent of their occurrences in this position were identified as "high." The result was a disproportionate number of "high" judgments for stimulus 3 and a disproportionate number of "low" judgments for stimulus 4. Note that, although precisely the same stimuli and test order were used for the consonant task, no discontinuity appears in the consonant function.

We defer discussion of these results until we have reported on the next experiment.

EXPERIMENT II

The purpose of this experiment was to explore further the effects of anchoring in correlated phonetic and nonphonetic continua, this time using intensity as the nonphonetic dimension varied over the stimulus continuum.

Method

Stimuli. The stimuli were the series of consonant-vowel syllables used in Experiment I (see Table 1), but with three diferences: (1) fundamental frequency for the first 225 msec was held constant at 114 Hz for all seven stimuli in the continuum and fell linearly to 80 Hz during the last 75 msec of each syllable; (2) in order to conform to most previous adaptation experiments, the stimuli were shortened to a duration of 250 msec by dropping 50 msec of the steady-state portion; (3) amplitude during the first 175 msec was attenuated by 18 dB in 3 dB steps from /bæ/ to /dæ/. Therefore, in this

experiment, the correlate of position on the phoneme continuum was overall amplitude.

Tapes were prepared in the same way as for Experiment I.

Procedure. Order of testing followed the same pattern as in Experiment I: loudness control, loudness anchor, phoneme control, phoneme anchor. In the loudness condition, subjects were asked to listen for loudness differences in the syllables and to ignore other differences observed. Instructions were to write "L" for a loud sound or "S" for a soft sound for the loudness judgments, and "b" or "d" for the phoneme judgments. Each experimental session was preceded by a display and a practice tape. The gain of the tape recorder playback for the 1000 Hz calibration tone was adjusted to give a voltage reading across the earphones equivalent to approximately 90 dB re 0.0002 dyne/cm^2. Therefore, the test stimuli varied from 90 dB SPL for stimulus 1 (the anchor stimulus) to 72 dB SPL for stimulus 7.

Nine paid listeners served as subjects. All met the criteria established in the previous experiment. They were tested either singly or in pairs, and testing was completed in one session with a short break between the loudness and phoneme judgment conditions.

Results

One subject was dropped from the data analysis because he did not categorize the stimuli consistently in the phoneme identification task. Figure 2 displays the loudness and consonant group identification functions for the remaining eight subjects in the control and anchor conditions. The slopes of the control functions seem equally steep for the two tasks, but the anchor function is appreciably lowered for loudness, very little for the consonants.

Table 5 presents the total number of "loud" and "b" responses (excluding those to the anchor stimulus) for each of the eight subjects under both conditions. For the loudness task, there is a significant mean decrease of 5.4 "loud" responses in the anchoring condition (matched pairs $t = 4.80$, $p < .005$, one-tailed). For the speech task there is an insignificant decrease of 1.2 "b" responses in the anchoring condition (matched pairs $t = 1.2$, $p > .10$, one-tailed).

Normal ogives were again fitted to the individual data, and the results are listed in Table 6. For the loudness task, the mean control and anchor boundaries are 3.78 and 3.41 respectively, a significant shift in the anchor condition of 0.37 continuum steps toward the anchor stimulus (matched pairs $t = 4.17$, $p < .005$, one-tailed). For the speech task, the mean control and anchor boundaries are 3.59 and 3.45 respectively, an insignificant shift in the anchor condition of 0.14 continuum steps toward the anchor stimulus (matched pairs $t = .09$, $p > .45$, one-tailed.

Table 7 lists individual and mean standard deviations of the fitted ogives. The mean loudness standard deviation of 1.07 is insignificantly larger than the mean speech standard deviation of 1.05 on a two-tailed matched pairs t-test ($t = 0.82$, $p > .40$). There is no significant rank order

correlation between control standard deviation and boundary shift for either loudness (rho = .30) or speech (rho = .22).

Discussion of Experiments I and II

The overall conclusion to be drawn from these experiments is essentially similar to that of Sawusch and Pisoni (1973) and of Sawusch et al. (1974): a synthetic stop consonant continuum is less susceptible to psychophysical anchoring effects than an arbitrary continuum of pitch or loudness. However, the conclusion is even stronger in the present experiments, since each value of place was always paired with a given F_0 or intensity value, so that one dimension was fully predictable from the other. Despite this deliberately imposed association of stimulus dimensions, phonetic and nonphonetic response patterns displayed definite dissociation.

One possible account of this dissociation is suggested by the fact that the stimuli most susceptible to anchoring effects are usually those for which the variance of judgment is highest (Volkmann, 1951; Parducci, 1965).[3] Certainly, the slopes of the pitch control functions in Experiment I were significantly greater than those of the speech control functions. Furthermore, all anchoring effects in both experiments (apart from the small effect on stimulus 2 in the loudness task) were confined to "boundary" stimuli that were not consistently identified in the control conditions. However, this cannot be the complete explanation for the dissociation, since there was no significant rank order correlation for either task in either experiment between the standard deviation of the control function and the degree of boundary shift; nor was there any significant difference in the slopes of the loudness and speech control functions of Experiment II. This last fact demonstrates that, even if members of a nonphonetic continuum are classified as consistently as members of a stop consonant continuum, they still may be more susceptible to anchoring. In other words, while some degree of stimulus ambiguity may be a necessary condition of anchoring effects, ambiguity alone cannot account for all of the variance. We return to this matter in the general discussion.

EXPERIMENT III

The results of the first two experiments may be taken as instances of a general susceptibility to context effects, typical of continuously perceived dimensions, as compared with a general resistance to context effects, typical of categorically perceived dimensions. Since strong context effects have been reported for synthetic vowels (Fry, Abramson, Eimas and Liberman, 1962; Stevens, Liberman, Studdert-Kennedy and Öhman, 1969) but less for synthetic stop consonants (Liberman, Harris, Kinney and Lane, 1961; Eimas, 1963), the next experiment extends the anchoring paradigm to compare its effects on correlated stop consonant and vowel continua.

---

[3]We are indebted to Katherine Harris for emphasizing this fact and for drawing our attention to the work of Volkmann.

Method

Stimuli. The stimuli in this experiment were a series of two-formant consonant-vowel syllables ranging perceptually from /bæ/ to /dɛ/ in approximately equal steps of second-formant transition onsets and first- and second-formant steady-state frequencies. The third-formant circuit on the Haskins parallel resonance synthesizer was turned off during synthesis because of the difficulty encountered in generating a perceptually acceptable three-formant correlated vowel and consonant continuum.

The stimulus values, including transition durations and extents, are displayed in Table 8. All F_1 transitions were 40 msec in duration and rose linearly from 234 Hz to a steady-state frequency ranging from 718 Hz (stimulus 1) to 562 Hz (stimulus 7), so that the transition extents ranged from 484 Hz to 328 Hz. The F_2 transitions rose linearly from onset frequencies ranging between 1385 Hz (stimulus 1) and 1845 Hz (stimulus 7) to steady-state frequencies ranging between 1541 Hz and 1996 Hz, so that transition extents remained roughly constant around 152 Hz. The rising F_2 transitions were necessary to maintain roughly equal frequency steps along the continuum. To counter the resulting labial percepts, F_2 transitions were reduced in duration from 40 msec (stimulus 1) to 20 msec (stimulus 7). The effect of this maneuver was presumably to reduce the perceptual salience of the transitions and thus "flatten" the second formants, granting the more abbreviated transitions (stimuli 5, 6, and 7) effective values closer to the more or less level transitions appropriate for /d/ in these vowel contexts. A pilot experiment established the acceptability of the stimulus series and estimated the phoneme boundaries. For each stimulus, fundamental frequency was constant at 114 Hz for the first 175 msec and dropped linearly to 80 Hz, while overall amplitude dropped by 28 dB, over the last 75 msec.

All experimental tapes were produced as previously described. Two test tapes were again prepared: a seventy-item control tape with all stimuli presented equally often, and a 100-item anchor tape with stimulus 1 (/bæ/) presented four times as often as stimuli 2 through 7.

Procedure. As in previous experiments, the study called for identification of items on the two test tapes under two different sets of instructions. In the first condition, the task was to categorize the vowels as /æ/ or /ɛ/; in the second condition the task was to identify the consonants as /b/ or /d/. Since we feared that the obvious correlation between vowel and consonant might tempt listeners to base their responses in both conditions on the same dimension, we decided to enhance the correlation by drawing subjects' attention to it, while asking them to disregard it in making their judgments. The order of testing was the same for all subjects: vowel control, vowel anchor, stop-consonant control and stop-consonant anchor. All other aspects of the procedure were identical to those of the previous experiments. There were nine paid listeners, all meeting the earlier criteria.

The experimental tapes were reproduced by an Ampex AC 500 tape recorder and presented binaurally through Telephonics (TDH-39) matched and calibrated headphones. The voltage across the headphones of the 1000 Hz calibration tone was set to the equivalent of approximately 80 dB re 0.0002 dyne/cm^2.

Results

One subject was dropped from the data analysis because she did not categorize the stimuli consistently in the consonant identification task. Figure 3 displays the vowel and consonant group identification functions for the remaining eight subjects in the control and anchor conditions. The slopes of the consonant functions are notably less steep over the lower end of the continuum than are those of the vowel functions. The anchor functions are appreciably lower than the control functions for both consonants and vowels.

Table 9 presents the total number of "æ" and "b" responses (excluding those to the anchor stimulus) for each of the eight subjects under both conditions. For the vowel task, there is a significant mean decrease of 6.7 "æ" responses in the anchor condition (matched pairs $t = 4.60$, $p < .005$, one-tailed). For the consonant task, there is a significant mean decrease of 3.5 "b" responses in the anchor condition (matched pairs $t = 2.4$, $p < .025$, one-tailed).

Normal ogives were fitted to the individual data, and the results are listed in Table 10. For the vowel task, the mean control and anchor boundaries are 3.87 and 3.46 respectively, a significant shift in the anchor condition of 0.41 continuum steps toward the anchor stimulus (matched pairs $t = 4.7$, $p < .005$, one-tailed). For the consonant task, the mean control and anchor boundaries are 3.62 and 3.32 respectively, a marginally significant shift in the anchor condition of 0.30 continuum steps toward the anchor stimulus (matched pairs $t = 1.51$, $p < .10$, one-tailed). Based on the results of Experiments I and II and of previous work with consonants and vowels, we may predict that any shift in the consonant boundary due to contextual anchoring effects will be significantly less than the shift in the vowel boundary on a correlated continuum, and this proves to be the case (matched pairs $t = 2.02$, $p < .05$, one-tailed).

Finally, the greater ambiguity of the consonants than of the vowels at the /bæ/ end of the continuum (Figure 3) was not evidenced by every subject. As may be seen from Table 11, which lists individual and mean standard deviations of the fitted ogives, only subjects 1, 4, 7, and 8 show a larger consonant than vowel standard deviation. For the other four subjects, consonant values are equal to or less than those for the vowels. The difference between the mean vowel standard deviation of 1.06 and the mean consonant standard deviation of 1.23 is marginally significant ($t = 1.99$, $p < .10$, two-tailed). There is no significant rank order correlation between control standard deviation and boundary shift for either vowels (rho = -.15) or consonants (rho = .02).

Discussion

Members of a synthetic stop consonant continuum are not always immune to the psychophysical effects of anchoring. The present effects are not strong, for though clearly significant ($p < .025$) when measured by the number of responses in the anchor response class, they are only marginally significant ($p < .10$) when measured by the boundary shift estimated from fitted normal ogives. Nonetheless, even on the latter measure, six of the eight subjects display the effect, and the mean boundary shift of 0.30 continuum steps is more than double that observed on the (differently constructed) continuum of Experiment II.

If we take these results to be reliable, two questions arise. First, why were synthetic stop consonants varying in the acoustic cues to place of articulation susceptible to anchoring along a correlated vowel continuum, but not along correlated continua of fundamental frequency or intensity? Second, why were such consonants less susceptible to anchoring than their correlated vowels?

We can immediately dismiss one possible answer to the first question. This is the suggestion that, having had their attention drawn to the consonant-vowel correlation, subjects tended to judge the vowel rather than the consonant, so that consonant judgments "followed" vowel judgments. If this were the case, we would expect some correlation both between consonant and vowel boundaries and between their boundary shifts. However, a scan of Table 10 suggests no such relations, and the Spearman coefficients of rank order correlation are not significant either for the control boundaries (rho = -.05) or for the boundary shifts (rho = .43).

A second possible answer is that the synthetic stops on the correlated vowel continuum were, at least toward the /bæ/ end of the continuum, poor exemplars of a phonetic class and therefore susceptible to contrast effects induced by anchoring. To the extent that ambiguous stimuli are more susceptible to anchoring than unambiguous stimuli, this interpretation would seem to be correct. We must then ask why these syllables were more ambiguous than those of Experiments I and II.

Here, we may recall that, since the correlated consonant-vowel continuum lacked a third formant, variations in place of articulation were cued entirely by second formant transitions. Furthermore, the onsets of these transitions were not adjusted to their following steady-state frequencies. While F_1 and F_2 steady-state frequencies varied to yield a range of vowels from /æ/ to /ɛ/, F_1 and F_2 onset frequencies were identical to those of Experiments I and II where steady-state formant frequencies were constant at values for /æ/. The onset frequencies (and transitions) were therefore somewhat inappropriate to the following vowels. If we add to this the fact that the vowel itself was unpredictable from trial to trial, so that the listener lacked the invariant spectral reference against which the variations of a synthetic stop consonant continuum are usually judged, it is not surprising that listeners found the continuum of Experiment III somewhat ambiguous.

To this uncertainty the /bæ/ anchor perhaps brought a note of stability. The recurrence of its relatively low F_2 transitional and steady-state frequencies may have established a standard against which the transitions and steady-states of the boundary stimuli (2, 3, and 4) were sometimes heard as raised and "flattened," the vowel more like /ɛ/, the consonant more like /d/. This explanation, in terms of psychophysical contrast, must, of course, be distinguished from the usual account in terms of detector "fatigue." A similar contrast explanation was offered by Blumstein, Stevens and Nigro (1977) for the adapting effect of a voiced velar stop consonant, cued by formant transitions without plosion.

The answer to our first question follows naturally from this account. Synthetic stop consonants were susceptible to anchoring on the correlated vowel continuum because simultaneous variations in transitional and steady-

state formant frequencies produced ambiguous stimuli that contrasted with the anchor stimulus over their entire length. The syllables of the correlated F_0 and intensity continua, on the other hand, were relatively unambiguous and contrasted with the anchor stimulus only in their first 40 msec. As we will suggest below in the general discussion, the total relevant energy in the anchoring stimulus and its accumulation over the test may largely determine its effect. It is precisely this which may also answer our second question and explain why the consonants of Experiment III were, despite their ambiguity, less susceptible to anchoring than their correlated vowels. This again is a matter to which we return in the general discussion.

EXPERIMENT IV

The preceding experiment demonstrated that, under certain conditions, an anchoring effect, similar to that of the standard adaptation paradigm, can be induced on a synthetic stop consonant continuum. In our final experiment, we undertook a direct comparison of the two paradigms on the covarying consonant and F_0 continua of Experiment I.

The defining distinction between the paradigms is in the distribution of anchor/adaptor energy over the test: the adaptation paradigm rapidly repeats the adaptors in a number of blocks immediately before the test stimuli, while the anchoring paradigm distributes its repetitions of the anchor stimulus randomly among the test syllables. A second distinction is in the ratio of anchor/adaptor stimuli to test stimuli: typically, the adaptor ratio is much the larger, often by a factor of several hundreds, if not thousands. (If the anchor ratio were appreciably increased, the effective differences between the two paradigms would presumably disappear.) A third difference between the paradigms is in the interstimulus intervals (ISI): while repetitions of the anchor stimulus, even if in immediate succession, are separated by the two or more seconds needed for a subject to respond, repetitions of the adaptor stimulus typically follow at a rate of one or two per second. A final, less important difference is that subjects are usually asked to respond to every stimulus presentation in the anchoring paradigm, but only to test syllables in the adaptation paradigm.

Taken together, these distinctions add up to a simple difference between the paradigms in the accumulated anchor/adaptor energy over the course of a test, a difference of degree rather than of kind. Since, further, adaptation effects have already been shown to depend on the number (Hillenbrand, 1975; Simon, 1977), intensity (Hillenbrand, 1975; Sawusch, 1977) and ISI (Simon, 1977) of the adaptors, we hoped to demonstrate by judicious juggling of these variables the essential equivalence of the two paradigms.

As a preliminary, it was established in two other experiments (Simon, 1977) that significant adaptation effects could be induced for synthetic stop consonants at a 16:1 adaptor-to-test stimulus ratio and at an ISI of 1750 msec. We took the 16:1 ratio to be close to the maximum at which the anchor and adaptation paradigms could still be effectively distinguished. We took the 1750 msec ISI to be close to the minimum within which subjects could be expected to respond. With these values of ISI and anchor/adaptor ratio, we were in a position to construct two tests, identical in every respect except for the placement of the anchor/adaptor syllables.

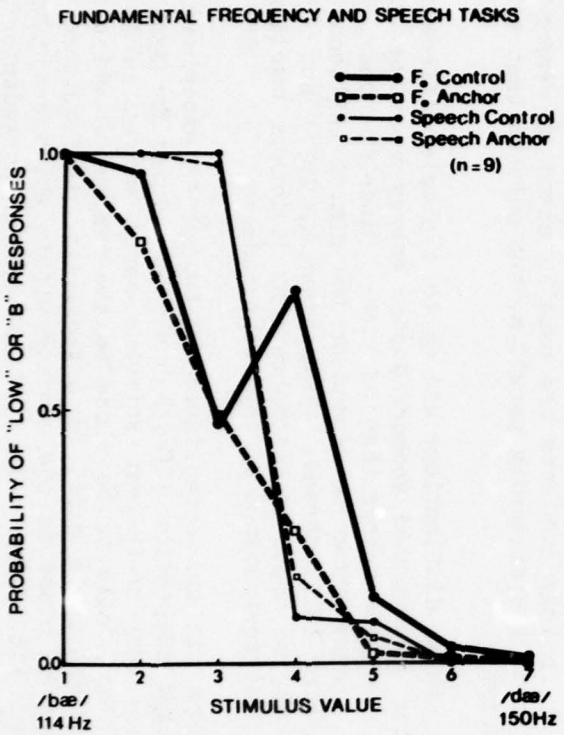


Figure 1: The probability of "low" or "b" responses as a function of stimulus value, under control and anchor conditions, for the pitch and speech tasks in Experiment I.

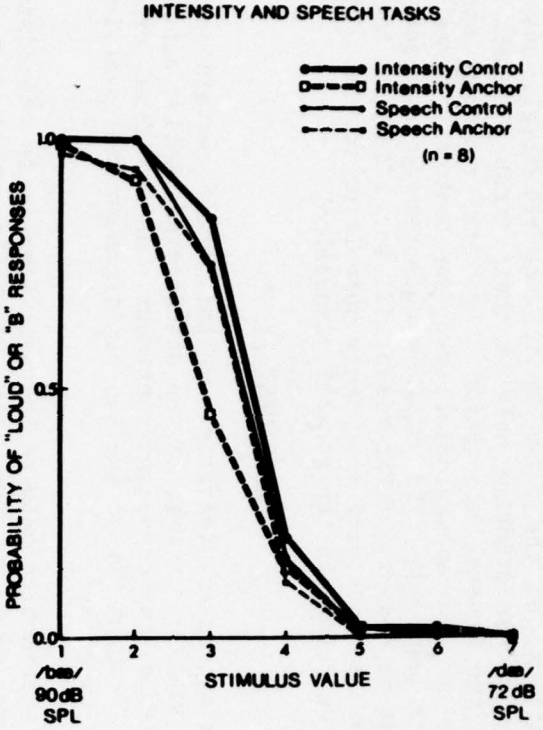


Figure 2: The probability of "loud" or "b" responses as a function of stimulus value, under control and anchor conditions, for the loudness and speech tasks in Experiment II.

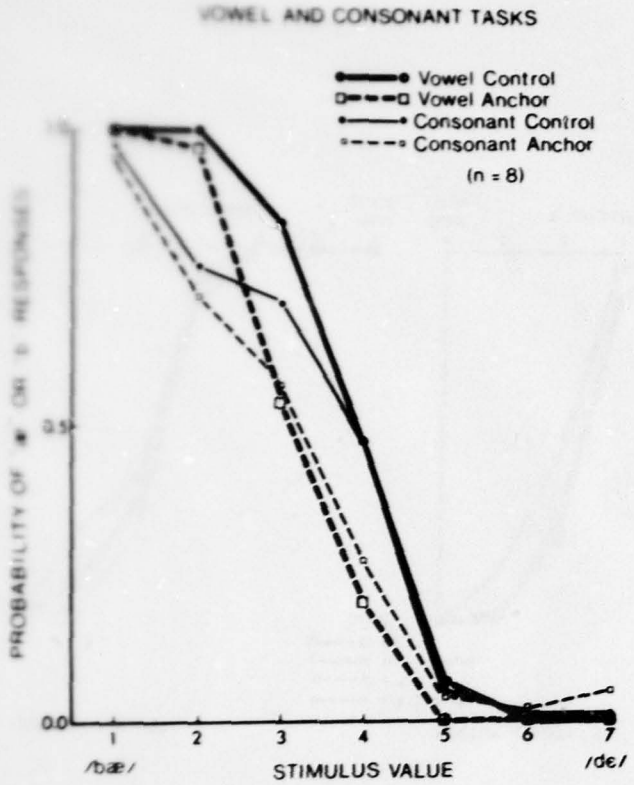


Figure 3: The probability of "ae" or "b" responses as a function of stimulus value, under control and anchor conditions, for the vowel and consonant task in Experiment III.

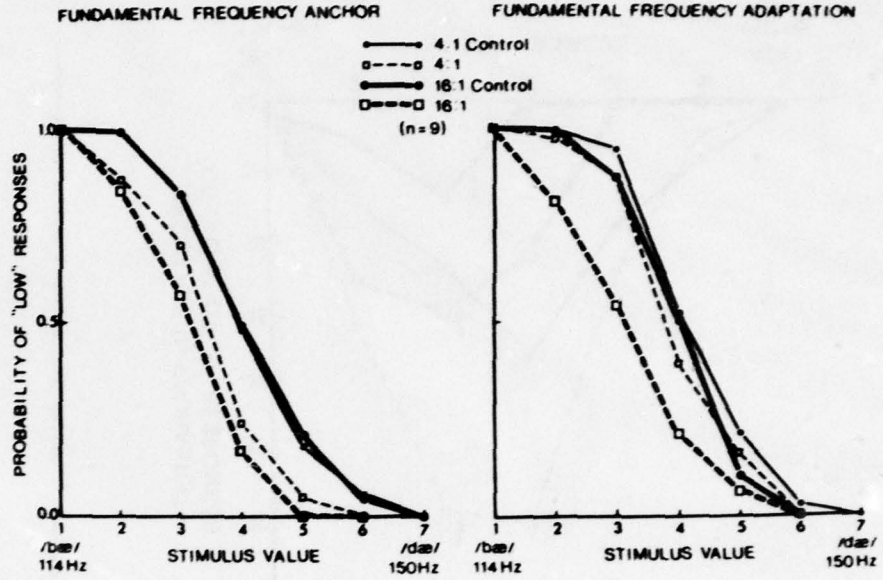


Figure 4: The probability of "low" responses as a function of stimulus value, in the two control and two ratio conditions of Experiment IV, for the pitch task, under anchoring (left) and adaptation (right).

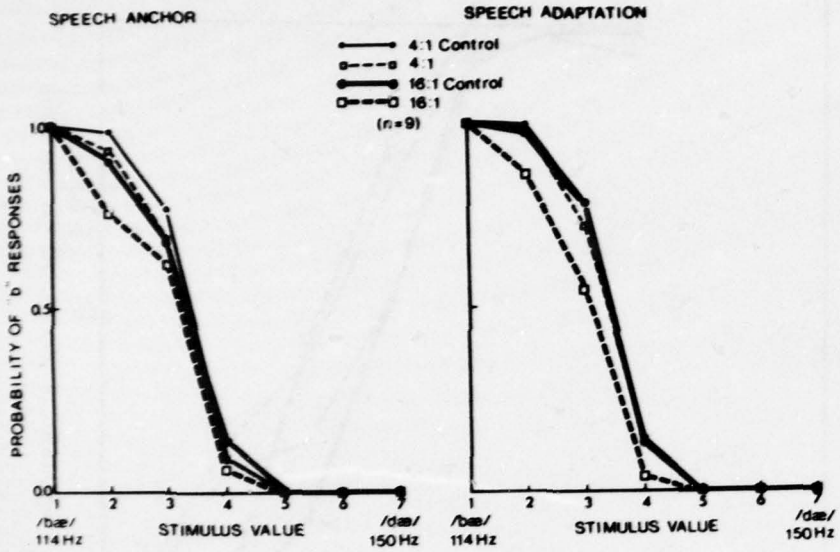


Figure 5: The probability of "b" responses as a function of stimulus value, in the two control and two ratio conditions of Experiment IV, for the speech task, under anchoring (left) and adaptation (right).

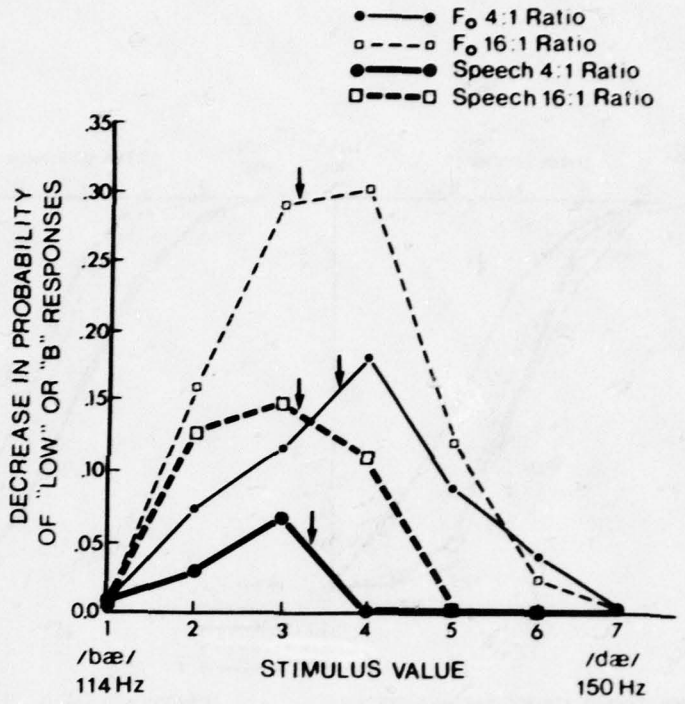


Figure 6: Mean decrease in the probability of "low" or "b" responses as a function of stimulus value, combined over anchoring and adaptation conditions, for the pitch and speech tasks in Experiment IV.

TABLE 1: Fundamental frequency and the starting frequencies of the second- and third-formant transitions for the synthetic CV syllables of Experiment I.

Stimulus Number	(Frequency in Hz) F_0	F_2	F_3
1	114	1385	2525
2	120	1468	2694
3	126	1541	2862
4	132	1620	3026
5	138	1695	3195
6	144	1772	3363
7	150	1845	3530

Note: The fixed steady state formants were centered at 666 Hz (F_1), 1620 (F_2) and 3026 (F_3).

TABLE 2: Individual and mean number "low" and "b" responses to the sixty presentations of stimuli 2 through 7 in Experiment I.

	F_0		SPEECH	
Subject	Control	Anchor	Control	Anchor
1	28	22	20	20
2	22	18	24	24
3	15	14	20	21
4	22	15	20	19
5	20	16	22	20
6	21	18	28	24
7	25	17	20	22
8	21	11	20	23
9	26	24	21	24
MEAN	23.3	16.1	21.7	21.9

TABLE 3: Individual and mean boundaries (means of the normal ogives for the F_0 and speech tasks in Experiment I.

	F_0		SPEECH	
Subject	Control	Anchor	Control	Anchor
1	3.28	3.87	3.67	3.67
2	3.72	3.42	3.92	3.96
3	3.15	2.93	3.67	3.80
4	3.68	3.23	3.67	3.50
5	3.60	2.09	3.87	3.97
6	3.79	3.39	3.72	3.92
7	3.63	3.48	3.67	3.88
8	3.83	2.85	3.67	3.92
9	4.04	3.91	3.81	3.90
MEAN	3.63	3.24	3.74	3.83

TABLE 4: Individual and mean standard deviations of the normal ogives fitted to the control identification data for the F_0 and speech tasks in Experiment I.

Subject	F_0	Speech
1	1.13	1.00
2	1.08	1.04
3	1.27	1.00
4	1.09	1.00
5	1.20	1.06
6	1.15	1.07
7	1.54	1.00
8	.93	1.00
9	1.08	1.04
MEAN	1.16	1.02

TABLE 5: Individual and mean number "loud" and "b" responses to the sixty presentations of stimuli 2 through 7 in Experiment II.

	Intensity		Speech	
Subject	Control	Anchor	Control	Anchor
1	17	10	20	19
2	19	12	24	23
3	21	9	13	8
4	21	18	23	17
5	17	13	20	18
6	32	28	19	26
7	19	16	14	15
8	21	18	20	17
MEAN	20.9	15.5	19.1	17.9

TABLE 6: Individual and mean boundaries for the intensity and speech tasks in Experiment II.

	Intensity		Speech	
Subject	Control	Anchor	Control	Anchor
1	3.36	2.72	3.40	3.84
2	3.58	3.12	3.96	3.92
3	3.80	3.10	3.54	2.22
4	3.82	3.68	3.78	3.36
5	3.48	3.54	3.66	3.38
6	4.64	4.26	3.50	4.04
7	3.58	3.32	3.22	3.28
8	4.00	3.56	3.64	3.58
MEAN	3.78	3.41	3.59	3.45

TABLE 7: Individual and mean standard deviations of the normal ogives fitted to the control identification data for the intensity and speech tasks in Experiment II.

Subject	Intensity	Speech
1	1.08	1.06
2	1.06	1.00
3	1.10	1.12
4	1.00	1.06
5	1.08	1.00
6	1.18	1.04
7	1.06	1.10
8	1.00	1.06
MEAN	1.07	1.05

TABLE 8: Onset and steady-state frequencies, extent of transition rises and transition durations for the two formants of the correlated vowel and consonant continuum (Experiment III). All syllables were 250 msec in duration.

	F_1				F_2			
Stimulus Number	Onset frequency in Hz	Steady-state frequency in Hz	Transition extent in Hz	Transition duration in msec	Onset frequency in Hz	Steady-state frequency in Hz	Transition extent in Hz	Transition duration in msec
1	234	718	484	40	1385	1541	156	40
2	234	692	458	40	1468	1620	152	40
3	234	666	432	40	1541	1695	154	35
4	234	640	406	40	1620	1772	152	30
5	234	614	380	40	1695	1845	150	25
6	234	588	354	40	1772	1920	148	25
7	234	562	328	40	1845	1996	151	20

TABLE 9: Individual and mean number "æ" and "ɔ" responses to the sixty presentations of stimuli 2 through 7 in the control and anchor conditions of Experiment III.

	Vowel		Consonant	
Subject	Control	Anchor	Control	Anchor
1	26	25	24	25
2	21	14	27	24
3	18	16	27	25
4	28	16	8	3
5	27	19	30	24
6	19	15	29	18
7	24	17	6	8
8	27	15	12	8
MEAN	23.8	17.1	20.4	16.9

TABLE 10: Individual and mean boundaries for the vowel and consonant tasks in Experiment III.

	Vowel		Consonant	
Subject	Control	Anchor	Control	Anchor
1	4.18	3.88	4.22	4.18
2	3.68	3.34	4.12	3.96
3	3.54	3.52	4.08	4.36
4	3.88	3.32	2.60	1.26
5	3.96	3.60	4.34	4.06
6	3.60	3.28	4.28	3.42
7	3.96	3.48	2.36	2.68
8	4.12	3.24	2.96	2.62
MEAN	3.87	3.46	3.62	3.32

TABLE 11: Individual and mean standard deviations of the normal ogives fitted to the control identification data for the vowel and consonant tasks in Experiment III.

Subject	Vowel	Consonant
1	1.06	1.38
2	1.08	1.04
3	1.08	1.00
4	1.00	1.40
5	1.04	1.04
6	1.08	1.04
7	1.00	1.56
8	1.12	1.34
MEAN	1.06	1.23

TABLE 12: Individual and mean numbers of "low" responses to the sixty presentations of stimuli 2 through 7 for the two ratio conditions and their associated controls under the anchoring and adaptation procedures of Experiment IV.

	Anchoring				Adaptation			
Subject	Control	4:1	Control	16:1	Control	4:1	Control	16:1
1	22	17	23	9	25	27	25	17
2	16	11	19	9	26	26	23	5
3	23	19	23	10	23	19	25	17
4	17	12	27	16	28	20	19	5
5	25	19	21	16	24	20	20	16
6	26	21	29	22	23	22	25	27
7	32	18	27	14	32	29	27	13
8	38	32	36	27	34	29	35	30
9	28	19	26	20	26	18	22	16
Mean	25.2	19.7	25.7	15.9	26.8	23.3	24.6	16.2

TABLE 13: Individual and mean boundaries on the F_0 functions for the two ratio conditions and their associated controls under the anchoring and adaptation procedures of Experiment IV.

	Anchoring				Adaptation			
Subject	Control	4:1	Control	16:1	Control	4:1	Control	16:1
1	3.74	3.28	3.92	2.62	4.04	4.19	3.85	3.50
2	3.36	2.77	3.71	2.74	3.95	3.84	3.87	2.22
3	3.81	3.49	3.92	2.70	3.30	3.58	3.81	3.31
4	3.50	2.96	4.11	3.20	4.31	3.58	3.45	2.22
5	4.00	3.58	3.72	3.32	3.94	3.65	3.65	3.14
6	3.93	3.80	4.20	3.77	3.36	3.77	4.00	4.12
7	4.62	3.21	4.11	3.03	4.45	4.14	4.22	3.00
8	5.04	4.45	5.03	4.08	4.74	4.18	4.73	3.99
9	4.2	3.43	4.19	3.64	4.07	3.55	3.77	3.43
Mean	4.03	3.44	4.10	3.23	4.02	3.83	3.93	3.21

TABLE 14: Individual and mean number of "b" responses to the sixty presentations of stimuli 2 through 7 for the two ratio conditions and their associated controls under the anchoring and adaptation procedures of Experiment IV.

	Anchoring				Adaptation			
Subject	Control	4:1	Control	16:1	Control	4:1	Control	16:1
1	13	8	10	3	12	11	14	9
2	21	20	18	18	21	23	20	20
3	20	20	21	16	21	20	21	17
4	17	14	19	20	19	18	17	14
5	25	26	27	22	23	23	24	24
6	19	20	14	20	14	13	20	11
7	20	15	18	6	22	19	20	11
8	9	11	4	2	16	14	9	8
9	21	21	24	22	24	25	24	15
Mean	18.3	17.2	17.2	14.3	19.1	18.4	18.8	14.3

TABLE 15: Individual and mean boundaries on the speech functions for the two ratio conditions and their associated controls under the anchoring and adaptation procedures of Experiment IV.

	Anchoring				Adaptation			
Subject	Control	4:1	Control	16:1	Control	4:1	Control	16:1
1	3.17	2.57	2.72	1.97	2.92	2.79	3.03	2.62
2	3.82	3.67	3.42	3.52	3.82	3.92	3.67	3.67
3	3.67	3.67	3.82	3.14	3.82	3.67	3.72	3.37
4	3.17	3.06	3.50	3.67	3.58	3.53	3.37	3.03
5	4.00	3.77	4.07	3.87	3.92	3.92	3.96	3.96
6	3.50	3.67	3.27	3.67	3.22	3.17	3.65	2.77
7	3.65	3.11	3.42	2.32	3.88	3.82	3.65	2.77
8	2.61	3.02	1.89	1.63	3.44	3.23	2.61	2.95
9	3.82	3.72	3.84	3.88	3.96	4.00	3.96	3.28
Mean	3.49	3.36	3.33	3.07	3.62	3.56	3.51	3.16

TABLE 16: Observed decreases percentaged against possible decreases in probability of "low" or "b" responses for combined anchoring and adaptation conditions of Experiment IV at two ratios of anchor (or adaptor) to test stimulus.

	Percentage Decreases			
	Fundamental frequency ("low")		Speech ("b")	
Stimulus	4:1	16:1	4:1	16:1
2	6	16	3	13
3	9	35	9	20
4	36	62	-	57
5	45	81	-	-
6	100	100	-	-

TABLE 17: Individual and mean standard deviations of the normal ogives fitted to the control identification data for the F_0 and phoneme tasks of Experiment IV.

	F_0				PHONEME			
	Anchor		Adaptation		Anchor		Adaptation	
Subject	4:1	16:1	4:1	16:1	4:1	16:1	4:1	16:1
1	1.06	1.08	1.12	1.06	1.12	1.29	1.22	1.19
2	1.25	1.15	1.22	1.13	1.00	1.06	1.00	1.00
3	1.04	1.00	1.00	1.04	1.00	1.00	1.00	1.04
4	1.04	1.41	1.22	1.15	1.12	1.04	1.06	1.07
5	1.00	1.04	1.00	1.04	1.00	1.00	1.00	1.00
6	1.10	1.08	1.15	1.08	1.04	1.09	1.11	1.04
7	1.19	1.10	1.08	1.06	1.04	1.06	1.00	1.04
8	1.29	1.22	1.18	1.09	1.31	1.55	1.09	1.31
9	1.22	1.25	1.10	1.04	1.00	1.04	1.00	1.00
Mean	1.13	1.15	1.12	1.08	1.07	1.13	1.05	1.08
Grand Mean		1.12				1.08		

Method

Stimuli. The stimuli were the series of correlated F_0-consonant-vowel syllables used in Experiment I, reduced to 250 msec in duration by dropping the last 50 msec of the steady-state portion and making appropriate adjustments in the F_0 and overall amplitude contours.

Tests were prepared in the same way as for previous experiments. The control identification tests were the same for both anchoring and adaptation: 7 stimuli recorded 10 times each in a random order with 1750 msec between stimuli and a 5-sec pause after every 20th stimulus.

Two adaptation tests were prepared, with adaptor-to-test-stimulus ratios of 4:1 and 16:1. Each test consisted of 10 blocks of the specified number of adaptors (either 4 or 16) followed by one presentation of each of the remaining six syllables on the continuum, to make a total of either 100 (4:1) or 220 (16:1) stimuli, separated by 1750 msec within a block and by 5 seconds between blocks.

Two anchor tests were prepared, identical in every respect to the adaptation tests, except that the repetitions of the anchor stimulus were distributed randomly among the six test syllables of each block.

Procedure. The procedure was identical in both the adaptation and anchor experiments. Subjects were required to identify all stimuli: both anchors/adaptors and test syllables, under both control and experimental conditions for both tasks (pitch and phoneme) at each ratio (4:1 and 16:1). Each subject heard the same tapes for each of the two task conditions. Only the instructions to the subjects varied as the task changed: they were identical with those of Experiment I.

The adaptation tests were administered to all subjects before the anchoring tests, and the entire set of pitch tests was presented first, as in Experiment I. Each experimental test (4:1 or 16:1) was preceded by its own control; the order of the tapes was counterbalanced within each paradigm and each task condition.

All experimental tapes were reproduced binaurally from the output of an Ampex AG 500 tape recorder over calibrated Telephonics (TDH-39) matched headphones with a circumaural seal. The 1000-Hz calibration tone was set to deliver a voltage across the phones equivalent to approximately 75 dB SPL re 0.0002 dyne/cm^2.

Nine listeners meeting the previously stated criteria participated in the experiments. They were tested singly or in pairs, in a quiet room, with a minimum of 24 hours and a maximum of a week between conditions.

Results

Figure 4 displays the pitch group identification functions in the control and experimental conditions for anchoring (left) and adaptation (right). The experimental functions are appreciably lower than the control functions under both paradigms.

Table 12 presents the total number of "low" responses (excluding those to the anchor/adaptor stimulus) under the two control and the two ratio conditions for both paradigms. In the anchor condition, there is a significant mean decrease in "low" responses of 6.5 at the 4:1 ratio (matched pairs $t = 6.24$, $p < .0005$, one-tailed), and of 9.8 at the 16:1 ratio (matched pairs $t = 4.23$, $p < .005$, one-tailed). In the adaptation condition, there is a significant mean decrease in "low" responses of 3.5 at the 4:1 ratio (matched pairs $t = 3.09$, $p < .01$, one-tailed), and of 8.4 at the 16:1 ratio (matched pairs $t = 4.12$, $p < .005$, one-tailed).

Normal ogives were fitted to the pitch data, and the results are listed in Table 13. In the anchor condition, the mean 4:1 control and experimental boundaries are 4.03 and 3.44 respectively, a significant shift of 0.59 continuum steps toward the anchor stimulus (matched pairs $t = 4.84$, $p < .005$, one-tailed). The mean 16:1 control and experimental boundaries are 4.10 and 3.23 respectively, a significant shift of 0.87 continuum steps toward the anchor stimulus (matched pairs $t = 7.84$, $p < .0005$, one-tailed). In the adaptation condition, the mean 4:1 control and experimental boundaries are 4.02 and 3.83 respectively, a marginally significant shift of 0.19 steps (matched pairs $t = 1.46$, $p < .10$, one-tailed); the mean 16:1 control and experimental boundaries are 3.93 and 3.21 respectively, a significant shift of 0.71 steps (matched pairs $t = 3.85$, $p < .005$, one-tailed).

Figure 5 displays the group speech identification functions in the control and experimental conditions for anchoring (left) and adaptation (right). The experimental functions are clearly lower than the control functions at the 16:1, but scarcely at the 4:1 ratio, under both paradigms.

Table 14 presents the total number of "b" responses (excluding those to the anchor/adaptor stimulus) under the two control and the two experimental conditions for both paradigms. In the anchor condition, there is an insignificant mean decrease in "b" responses of 1.1 at the 4:1 ratio (matched pairs $t = 1.26$, $p > .10$, one-tailed), and a marginally significant mean decrease of 2.9 responses at the 16:1 ratio (matched pairs $t = 1.69$, $p < .10$, one-tailed). In the adaptation condition, there is a marginally significant mean decrease in "b" responses of 0.7 at the 4:1 ratio (matched pairs $t = 1.40$, $p < .10$, one-tailed), and a significant decrease of 4.5 responses at the 16:1 ratio (matched pairs $t = 3.54$, $p < .005$, one-tailed).

Normal ogives were fitted to the speech data, and the results are listed in Table 15. In the anchor condition the mean 4:1 control and experimental boundaries are 3.49 and 3.36 respectively, an insignificant shift of 0.13 continuum steps toward the anchor stimulus (matched pairs $t = 1.23$, $p > .10$, one-tailed). The mean 16:1 control and experimental boundaries are 3.33 and 3.07 respectively, a marginally significant shift of 0.26 continuum steps toward the anchor stimulus (matched pairs $t = 1.57$, $p < .10$, one-tailed). In the adaptation condition, the mean 4:1 control and experimental boundaries are 3.62 and 3.56 respectively, a significant shift of 0.06 steps (matched pairs $t = 1.90$, $p < .05$, one-tailed). The mean 16:1 control and experimental boundaries are 3.51 and 3.16 steps respectively, a significant shift of 0.35 steps (matched pairs $t = 2.52$, $p < .025$, one-tailed).

Two within-subjects analyses of variance were carried out. The first examined the differences between decreases in the number of "low" and "b" responses as a function of ratio, paradigm and task. The mean decrease in response due to the 16:1 ratio (6.3 responses, averaged across paradigms and tasks) was significantly greater than that due to the 4:1 ratio (3.0 responses) ($F = 35.28$; $df = 1,8$; $p < .01$). The mean decrease in responses on the pitch task (7.0 responses) was significantly larger than that on the speech task (2.3 responses) ($F = 19.42$; $df = 1,8$; $p < .01$). There was no significant effect of paradigm ($F = 1.36$; $df = 1,8$; $p > .25$) and there were no significant interactions.

The second analysis of variance examined the differences between boundary shifts as a function of ratio, paradigm and task. The mean boundary shift due to the 16:1 ratio (0.55 continuum steps, averaged across paradigms and tasks) was significantly greater than that due to the 4:1 ratio (0.24 steps) ($F = 12.11$; $df = 1,8$; $p < .01$). The mean boundary shift on the pitch task (0.59 steps) was significantly greater than that on the speech task (0.20 steps) ($F = 19.61$; $df = 1,8$; $p < .01$). There was no significant effect of paradigm ($F = 2.05$; $df = 1,8$; $p > .10$), and there were no significant interactions.

In order to illustrate how the effects of anchoring and adaptation are distributed over the stimulus series, Figure 6 replots the data of Figures 4 and 5 in a format borrowed from Bailey (1973, 1975). The mean decrease in the probability of "low" or "b" responses, due to anchoring or adaptation, is plotted as a function of stimulus value, with task and ratio as parameters of the curves. *Since there was no significant paradigm effect, anchoring and* adaptation are combined. The arrows over the curves mark the positions of the mean boundaries estimated for the control conditions. The response shifts are obviously very much larger for pitch than for the phoneme, but for both tasks the shifts are confined to stimuli that were not identified with total consistency in the control conditions (that is, Stimuli 2, 3 and 4 for the speech task and 2 through 6 for the pitch task). Notice that on every function the absolute response shift increases with distance from the adaptor to a peak near the category boundary and thereafter declines. The decline is a floor effect entailed by the dichotomous scale: if the observed decreases in response probability are computed as percentages of possible decreases, monotonic increasing functions over the entire region of ambiguity result (see Table 16).

Finally, Table 17 lists individual and mean standard deviations of the normal ogives fitted to the control identification data for the pitch and speech tasks under both paradigms. The grand mean standard deviation for pitch of 1.12 is slightly larger than the grand mean standard deviation for speech of 1.08, but one-way analysis of variance shows no significant effect of task ($F < 1$). The Spearman coefficients of rank order correlation between control standard deviation and boundary shift were significant for the pitch task under anchoring at 4:1 ($rho = .63$, $p < .05$) and under adaptation at 16:1 ($rho = .59$, $p < .05$), but not under anchoring at 16:1 ($rho = -.17$) or under adaptation at 4:1 ($rho = .33$). None of the coefficients was significant for the speech task under either anchoring (4:1, $rho = -.17$; 16:1, $rho = .10$) or adaptation (4:1, $rho = .48$; 16:1, $rho = -.37$).

Discussion

The effect of paradigm was not significant by analysis of variance and did not interact significantly with either task or ratio. Since, further, the effect of ratio was significant, without interactions, we may conclude that selective adaptation and psychophysical anchoring are equivalent procedures having similar effects on pitch and stop consonant identification, at least under the conditions of this experiment. The conclusion is weak, since the effect of anchoring on the consonant task was certainly marginal. Furthermore, the effect of adaptation on both tasks was probably reduced by the use of low ratios and a relatively long interval between adaptor presentations. Yet, taken with the small but significant effect of anchoring on the consonants of Experiment III, the results do suggest that the two paradigms differ in degree rather than in kind.

The effect of task was clearly significant, again without interactions. This result confirms for anchoring, and extends to adaptation, the finding of Experiment I and of Sawusch, Pisoni and Cutting (1974) that the category boundary on an arbitrary F_0 continuum is significantly more susceptible to movement than is the category boundary on a stop consonant continuum. We will consider the possible origins of this difference more fully below. Here we note merely that, although the difference evidently reflects a differential influence of test context, it cannot be predicted simply from the ambiguity of the stimulus series, as measured by the standard deviations of the fitted control ogives, since these were neither significantly different between tasks nor reliably correlated with degree of boundary shift among subjects.

Finally, the lack of a paradigm effect, taken with the significant dissociation in anchoring effects between consonant and vowel tasks in Experiment III, and between consonant and pitch tasks in the present experiment, show that anchoring may be no less "selective" than adaptation: its effects are on perceptually distinct dimensions of the syllable rather than on the syllable as a whole.

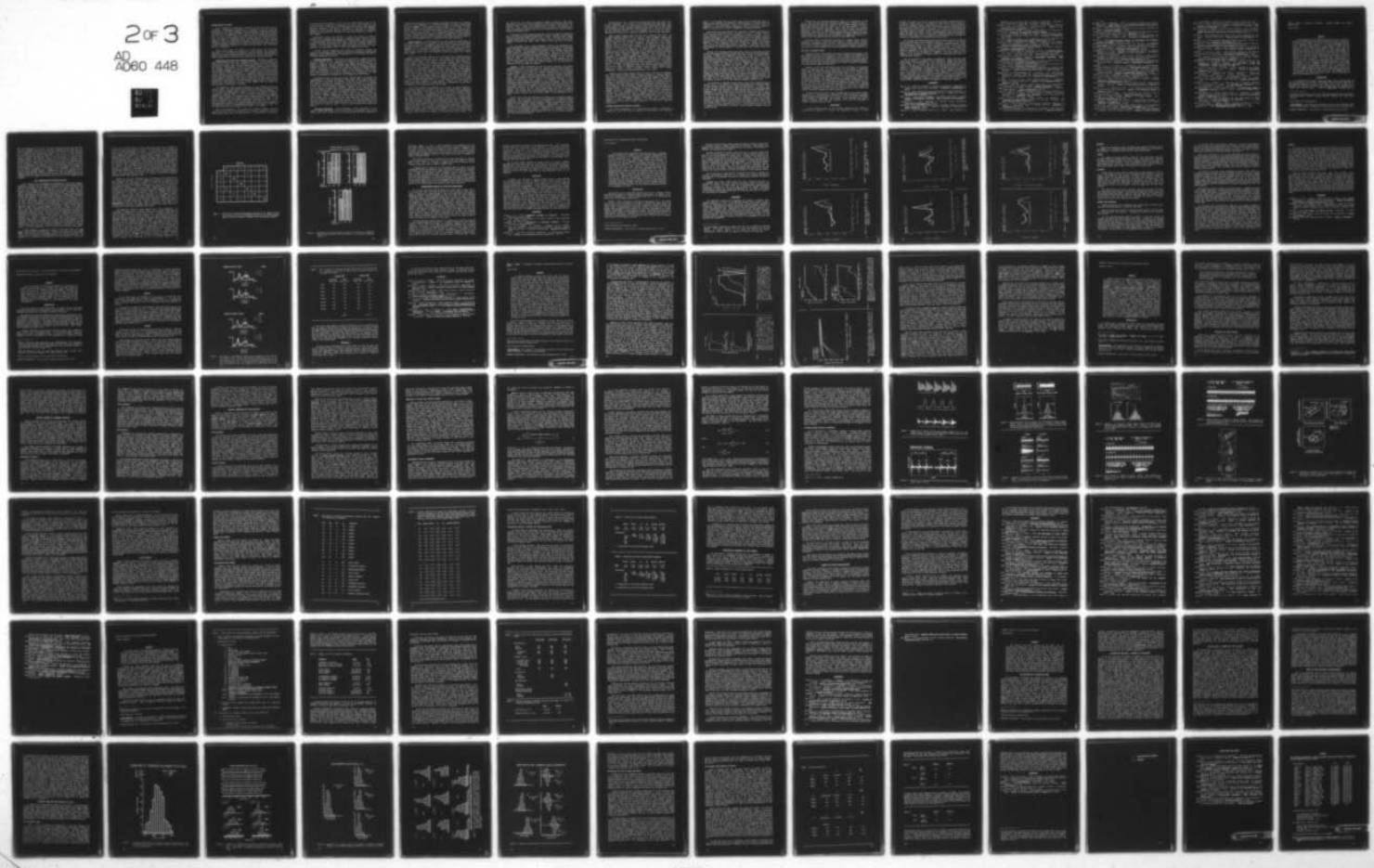
GENERAL DISCUSSION

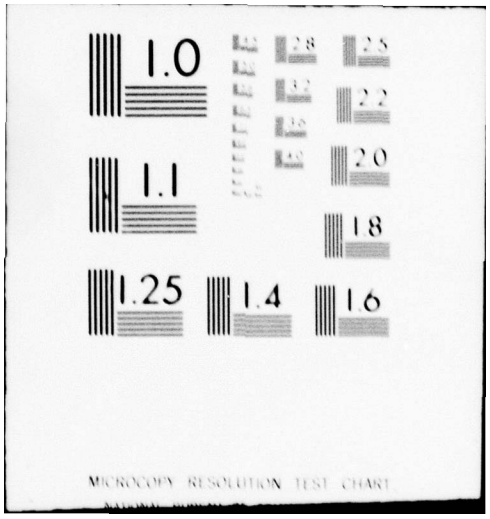
Task Differences

One unequivocal conclusion from these experiments is that a continuum of fundamental frequency is significantly more susceptible to the contextual effects of both anchoring and adaptation than a synthetic stop consonant place continuum, even when the two types of variation are carried, in perfect correlation, on the same syllables. If we may take the anchoring results of Experiments II and III and extrapolate from the lack of a paradigm effect in Experiment IV, both intensity and vowels are also more susceptible to anchoring and adaptation than stop consonants. The addition of vowels to the list forbids us to attribute the different susceptibilities to differences between phonetic and nonphonetic dimensions. However, we defer further discussion of the task differences until we have considered possible processes underlying anchoring and adaptation.

AD-A060 448 HASKINS LABS INC NEW HAVEN CONN F/G 5/7
SPEECH RESEARCH. A REPORT ON THE STATUS AND PROGRESS OF STUDIES--ETC(U)
JUN 78 A M LIBERMAN V101(134)P-342
UNCLASSIFIED SR-54(1978) NL

2 OF 3
AD A060 448




1.0
1.1
1.25
1.4
1.6
2.8
2.5
2.2
2.0
1.8
MICROCOPY RESOLUTION TEST CHART

Psychophysical Processes

If we reframe our unequivocal conclusion to make the point that a continuous, suprasegmental dimension of a syllable has proved even more susceptible to selective adaptation than a categorical, segmental dimension, the hypothesis that speech adaptation reflects fatigue of specialized feature detectors is not strengthened. For it was the facts of categorical perception that, at least in part, initially spurred the search for specialized detectors. Nor does the hypothesis gather strength from the finding that the selective effects of adaptation, which lent plausibility to the postulated fatigue of selectively tuned feature detectors, not only may result from the far weaker contextual manipulation of simple psychophysical anchoring, but may "select" fundamental frequency or intensity no less readily than consonantal place of articulation.

What, in fact, seems to be needed is a unified account that will handle both selective anchoring and adaptation of both phonetic and nonphonetic dimensions. In the following discussion we consider several standard psychophysical approaches, paying particular attention to how well they handle the speech adaptation data, for which the broader account seems most needed.

Adaptation Level Theory and Frequency Analysis. According to adaptation level theory (Helson, 1964, 1971), a subject, invited to classify members of a stimulus series into two or more categories and lacking any explicit standard with which to compare them, takes the psychological midpoint of the series as his standard. This midpoint, or point of subjective equality (usually termed the "category boundary" in speech adaptation studies), proves to be some appropriately weighted measure of the central tendency of the stimulus series. If the series is skewed by addition of an extreme stimulus at some distance from an endpoint or by presentation of a particular stimulus more often than others, the weighted mean of the series is shifted and the entire distribution of the subject's responses shifts also to match the new mean. Thus, in the anchor/adaptor conditions of the present experiments, the increased presentations of the lower endpoint of the series gave a negative skew to the stimulus distribution, reducing both the mean of the stimuli, and, in theory, the matching mean of the subject's responses.

An alternative, closely related formulation is the frequency analysis of Parducci (1965, 1974) who argues that the general tendency of subjects, faced with an absolute judgment task for which they lack an explicit standard, is "...to place the same number of stimuli in each of the available [response] categories" (Parducci, 1974, p. 134). If stimulus frequencies are unequal, subjects will distribute a frequent stimulus over several response categories and/or combine several rare stimuli into a single category. Thus, in the control conditions of the present experiments, subjects would tend to call half the stimuli "low" or "b" and the other half "high" or "d". In the anchor/adaptor conditions the increased presentations of the lower endpoint of the series would lead to a corresponding decrease in the number of times that the remaining stimuli were assigned to the "low" or "b" category.

While both adaptation level theory and frequency analysis correctly predict the direction of the changes in the subjects' responses, neither of them predicts their extent with any precision. In Experiment IV, for example,

the boundary shifts induced by the 4:1 and 16:1 ratios (Tables 13 and 15) are in the direction predicted by adaptation level theory, but clearly do not match the shifts in the mean of the stimulus series, whatever the weighting procedure. Similarly, the drops in the number of "low" and "b" responses induced by the two ratios (Tables 12 and 14) are predicted by a frequency analysis, but obviously fall far short of placing equal numbers of stimuli in each category.

Furthermore, since both theories "...treat the context or frame of reference as a frequency distribution of stimulus values" (Parducci, 1974, p. 128), they both predict that changes in that distribution will be reflected by response shifts across its entire range. However, most speech adaptation studies, including the present one, have found that response shifts are confined to a few boundary stimuli toward the adaptor end of the continuum. Certainly, none has shown a shift of the entire response distribution.

In short, neither adaptation level theory nor frequency analysis can provide a satisfactory account either of the effects induced by anchoring and adaptation in the present series of experiments or of the effects reported in the speech adaptation literature.

Response Organization. As we saw in the introduction, Sawusch and Pisoni (1973) and Sawusch, et al., (1974) took the immunity of consonant continua to anchoring effects in their experiments as evidence that consonants are not subject to response bias and inferred that speech adaptation effects are therefore perceptual. However, reliable range effects on consonant continua, probably attributable to response bias, have been reported by Studdert-Kennedy (1976b). Furthermore, if we accept either the contention of Helson (1971) that anchoring effects are, in fact, sensory, or the conclusion of the present study that anchoring and selective adaptation are essentially the same process, the arguments of Sawusch and his colleagues lose their force.

Nonetheless, there is other, more direct evidence against a response bias interpretation of selective adaptation [see Cooper (1975), Ades (1976) and Eimas and Miller (in press) for reviews]. First are the cross-series effects, the repeated finding that boundary shifts may be induced by an adapting stimulus not drawn from the test series. These effects rule out response bias at the level of the syllable or phoneme. Second are all those studies in which the degree of adaptation declines as spectral overlap between adaptor and test series declines. Third is the evidence that the degree of adaptation varies with the extent to which the adapting stimulus is a good exemplar of the adapted category (Sawusch and Pisoni, 1976; Miller, 1977a; Cole and Cooper, 1977). Finally, asymmetries in the effects of adapting stimuli drawn from opposite ends of a continuum have been repeatedly reported, since the initial observation of Eimas and Corbit (1973) that their voiceless adaptors were more effective than their voiced adaptors. The combined weight of the evidence does not preclude response bias as a factor (see Cooper, Ebert and Cole, 1976), but certainly rules it out as a major determinant of adaptation effects.

Perceptual Sharpening. A third hypothesis, briefly considered by Cole, Cooper, Singer and Allard (1975) and the focus of experiments by Cole and Cooper (1977) and by Ainsworth (1977), would attribute selective adaptation to

"retuning" or sharpening of the perceptual mechanism responsible for assigning stimuli to their categories: "...repeated listening to a single speech sound narrows the range of stimuli which are now acceptable as belonging to the same phonetic category as the adapting syllable" (Cole, et al., 1975, p. 243). Cole and Cooper (1977) falsified the hypothesis for speech by showing that a single sound was not required: they induced significant adaptation effects with a randomly repeated series of six syllables, spanning an entire phonetic category. Ainsworth (1977) reached a similar conclusion, and further grounds for rejecting this hypothesis come from any study demonstrating an effect of adaptation on the adaptor itself. Instances of this are cited below.

Anchor Contrast. A fourth hypothesis is that repeated presentations of the anchor/adaptor stimulus establish an auditory ground with which ambiguous test stimuli contrast. For example, adaptation with a stimulus having a short VOT makes the VOT of a midrange stimulus seem longer than it is. Or, as was suggested in our discussion of Experiment III, repeated presentation of a syllable with an initially rising F_2 may make the flat transition of a mid-range stimulus seem to fall (cf. Blumstein, Stevens and Nigro, 1977, Figure 13, p. 1312).

There is no doubt that contrast can be a potent force in shaping absolute judgments of vowel continua (Fry, et al., 1962; Eimas, 1963) and of arbitrarily labeled nonphonetic continua (Sherif, Taub and Hovland, 1958; Eimas, 1963; Helson, 1964). Recall, for example, the large contrast effect exerted on pitch judgments by immediately preceding stimuli in Experiment I. Of course, we may also recall, from the same experiment, the complete absence of an effect on the stop consonants. But this may be an instance of the general resistance of consonants to context effects (which we will discuss below) rather than of their total immunity. In fact, contrast effects on judgments of both voicing and place of articulation in stop consonants were reported by Eimas (1963). More recently, Brady and Darwin (in press) manipulated the range of synthetic VOT stimuli within an identification test block and found that "...the perceived voicing of a sound depends quite markedly on the range of other sounds presented before it.... The more voiced are the previous sounds, the more voiceless it will appear." Finally, our own place of articulation effects with a 4:1 anchoring ratio in Experiment III can hardly be explained by any mechanism other than contrast.

If we are to extend the hypothesis to an account of speech adaptation, the results summarized in Figure 6 and Table 16 are particularly important, because their form is typical of both speech adaptation and psychophysical contrast experiments. For example, Sherif et al., (1958) studied contrast effects in the rating of weights against a fixed standard on a 6-point scale. For stimuli close to the standard they found a modest shift in ratings toward the anchor value (assimilation), but beyond a certain point, ratings shifted away from the anchor (contrast) and the magnitude of the shift increased with the distance between the anchor and the stimulus being rated. Eimas (1963) reports the same increasing function in his analysis of contrast effects in ABX triads drawn from a variety of phonetic and nonphonetic, visual and auditory continua. In the typical speech adaptation study, with its endpoint adaptor and 2-point scale, assimilation of stimuli close to the adaptor could hardly be detected, even if it were occurring, since, already in the control condition, these stimuli are assigned to the same class as the adaptor. But

for stimuli in the region of ambiguity between categories, where most adaptation effects occur, speech adaptation studies have invariably found that response shifts increase with distance from the adaptor (see Figure 6 and Table 16). Thus, a function that characterizes the results of experiments in simple psychophysical contrast turns out to characterize the results of speech adaptation studies as well.

A corollary of this increasing function is that the degree of adaptation should decrease as the stimulus selected for use as an adaptor is moved inward along the continuum (compare Foreit, 1977, p. 347). Precisely this outcome (though usually taken to reflect the "tuning curve" of a detector) has been reported by Sawusch and Pisoni (1976), Miller (1977a) and Cole and Cooper (1977).

For the moment, then, we may reasonably entertain the hypothesis that speech adaptation effects on ambiguous stimuli in the middle of the stimulus range, the region to which most adaptation effects are confined, result from auditory contrast. The hypothesis is attractive, since it might comfortably gather speech and nonspeech under a single rubric. However, this cannot be the whole story because within-category adaptation effects also occur: as we shall see shortly, adaptation may even affect response to the adapting stimulus itself. This outcome could only result from contrast, if the contrast were between the adaptor and some property of the stimulus distribution, a possibility that we have already rejected in our discussion of adaptation level and frequency analysis.

Fatigue. A final hypothesis, and the one usually preferred in the speech adaptation literature, is that adaptation response shifts reflect fatigue, or desensitization, of a tuned detector. An obvious prediction from this model is that adaptation will shift the response probabilities for the adapting stimulus itself. Such an effect was, in fact, reported by Warren and Gregory (1958), using a quite different experimental paradigm ("verbal transformation"), later systematically elaborated by Goldstein and Lackner (1973) and by Lackner and Goldstein (1975). However, the standard procedure of most adaptation studies conceals the effect, if it is present, by using a closed response set and eschewing any estimate of sensitivity changes. This fact has not troubled feature detector theorists because the adaptor is typically close to the modal value of its category, a value to which the opponent detector is supposedly insensitive, so that, no matter how great the fatigue, the adapted detector will always have an output greater than that of its supposed opponent. Yet a direct test of the fatigue hypothesis is clearly crucial to feature detector accounts and has recently begun to draw experimental attention.

Miller (1975, 1977) has attacked the problem with an ingenious dichotic procedure. She has shown, for both labial/alveolar and voiced/voiceless oppositions, that adaptation with a good exemplar of a phonetic category reduces the effectiveness of that exemplar in dichotic competition with a good exemplar of an opponent phonetic category. Unfortunately, while this is precisely the outcome that a fatigue account predicts, it is also the outcome that the contrast hypothesis predicts: adaptation will increase the relative salience of the unadapted feature and therefore its probability of being correctly identified against the background of (that is, in dichotic competition with) the adapted feature.

However, Miller, et al. (1977) have recently reported a more direct test of the hypothesis. The details of the experiment are complicated, but, briefly, they find that, in a three-choice /b,d,g/ identification test, using a good exemplar of each phoneme, adaptation with either /bæ/ or /gæ/ raises the intensity required to maintain a given probability of correctly identifying the adapting syllable. To our knowledge, this is the only direct demonstration of the drop in sensitivity predicted by the fatigue hypothesis.

Ample indirect support comes from the work of Sawusch (1976, 1977). He increased response sensitivity, by using a rating procedure rather than straight identification on a /b-d/ continuum, and repeatedly demonstrated significant drops, after adaptation, in the rated quality of an endpoint adaptor itself and of its near neighbors. Interestingly, these within-category drops in rated quality occurred if the adaptor was a test series endpoint, but not if the three formants of the adapting syllable, though identical in structure to those of a test series endpoint, were raised above them by 1 - 1 1/2 critical bandwidths. In other words, a standard endpoint adaptor produced drops in the rated quality of stimuli both within and between categories, while a spectrally displaced adaptor produced rated quality drops only of stimuli between categories (Sawusch, 1977, Experiment II). Furthermore, in this same experiment, Sawusch found only 50 percent interaural transfer of the adaptation effect for the standard adaptors, but 100 percent transfer for the spectrally displaced adaptors. From these and other results he inferred two levels of processing in selective adaptation: "...a frequency specific, peripheral auditory level and a more abstract, central, integrative level..." (p. 748). He suggested further that within-category response shifts may reflect fatigue of peripheral feature detectors, while between-category shifts may reflect central processes, either fatigue or "...other mechanisms...possibly involving decision rules..." (p. 749).

Broadly, this account goes well with our earlier evidence and arguments. Apart from adaptation-level theory, frequency analysis or response organization, each of which we have already rejected for other reasons, there seems no alternative to "fatigue" as an account of the adapting effect of a stimulus on itself, and the data of Sawusch (1977) suggest that the account should be extended to adapting effects on close spectral neighbors. We need not concern ourselves at this point with just what is fatigued beyond remarking that, if the fatigue is indeed peripheral, "feature detectors" would not seem to be likely candidates. As for the second level, reflected by the between-category response shifts, our earlier arguments suggest that auditory contrast--a somewhat "abstract", presumably "central" and perhaps even "integrative" process--may be an important factor. However, other factors must also be involved. For example, Diehl (1975) and Sawusch and Pisoni (1976) have shown that the adapting effect of a syllable may be determined less by its acoustic structure or by experimental labeling instructions than by the phoneme category to which a listener assigns it. In other words, whatever role auditory contrast may play in speech adaptation, it plays within linguistic constraints.

Phonemic Anchoring and Auditory Contrast

The central assumption of all psychophysical accounts of anchoring or adaptation is that the frame of reference is established by the experiment

itself. It is assumed that the subject has no prior standard with which to compare items presented for judgment and therefore derives a standard from the conditions of the experiment. While this assumption may be valid for judgments of arbitrarily labeled dimensions, such as brightness, weight or pitch, it clearly does not hold for judgments of consonantal status.

Many studies have demonstrated that a synthetic consonant continuum has a relatively firm perceptual structure, brought to it by the listener from his native language (for a review see Strange and Jenkins, in press). A two-category continuum, for example, ranges from one or more acceptable tokens of a particular phoneme to one or more acceptable tokens of an opponent phoneme; these extremes sufficiently resemble naturally occurring types for them to be assigned to their categories with high consistency. Between the extremes occur one or more ambiguous stimuli of types that perhaps never occur naturally and may even be articulatorily impossible. These are the stimuli that succumb to varying degrees of contrast with other stimuli within the series.

Whether the surrounding fixed frame either limits or facilitates contrast between these ambiguous central stimuli and other portions of the continuum is not known. However, some suggestions come from a study by Donald (1976), independently replicated in all its essentials by Foreit (1977). As a test continuum Donald used a labial VOT series from -80 to +70 msec, a range that spans three phonological categories (with boundaries at roughly -20 and +25 msec) for speakers of Thai, but only two categories (with a boundary at roughly +15 msec) for speakers of English. Among Donald's adaptors were syllables with VOT values of -80 and +5 msec. The latter produced a significant shift in the +15/+25 msec boundary for both groups. A simple auditory contrast hypothesis would predict even greater shifts after adaptation with the more distant adaptor (-80 msec). In the event, the distant adaptor (-80 msec) produced the same shift as the near adaptor (+5 msec) for the English speakers who judged them to be tokens of the same phoneme, but for Thai speakers, who judged the distant adaptor to be a different phoneme than the near, and to be separated from the +25 msec boundary by an entire phonological category, the -80 msec adaptor produced no effect at all.

Are we forced to conclude, as does Foreit (1977) after reporting similar results, that "...the effects of acoustic manipulations on selective adaptation are strongly limited by their linguistic implications..." (Foreit, 1977, p. 351) and that auditory contrast plays no role at all in producing boundary shifts? Perhaps; but consider a curious finding from Cole and Cooper (1977, Experiment II). They constructed a four-item test series from /ǰa/ to /da/ by trimming the duration of the initial friction on a naturally spoken syllable (37, 29, 21 and 14 msec of friction). In a similar manner they constructed two adaptors, /čі/ and /ǰі/, with friction durations of 120 and 60 msec. Since the 60 msec adaptor (/ǰі/) produced a significant shift in the /ǰa-da/ boundary, we might reasonably predict either, by simple contrast theory, that the 120 msec adaptor (/čі/) would produce an even greater shift or, by analogy with the Thai results reported above, that it would produce none at all. As it happens, both predictions are wrong: the two adaptors produced identical boundary shifts.

Although the cross-language differences remain to be explained, we may now be less confident that an intervening phonological category protects a phoneme boundary from adaptation effects or that different adaptors have the same effect because they are judged to be the same phoneme. In fact, phonological status may be irrelevant. Sarris (1967) has shown that contrast effects do not increase indefinitely with distance between standard and comparison: there is a critical distance beyond which they begin to decline. The equivalence of the near and distant adaptors for English speakers in these three studies may therefore be the equivalence of symmetrical points on a parabola, with the peak of auditory contrast lying between them.

In short, while the perceived structure of a speech series probably influences the acoustic range over which auditory contrast occurs, we know essentially nothing about how it does so. Before we dismiss auditory contrast as a factor in speech adaptation, we need systematic parametric studies of contrast over acoustic ranges within which representative speech continua lie.

The Role of Stimulus Energy

We come finally to the question with which we began this discussion: Why were judgments of place of articulation in stop consonants less susceptible to anchoring or adaptation than judgments of pitch, loudness and vowels? We saw repeatedly in the reports of results that this fact could not be attributed to systematic differences in stimulus ambiguity as measured by judgment variability. Nor have we come upon any factor in our review of possible psychophysical processes that might be implicated.

Perhaps, in fact, the most likely source of the difference is accumulated anchor or adaptor energy. We already know that the degree of adaptation varies with the number of adaptors (Experiment IV; see also Hillenbrand, 1975; Simon, 1977) and with the interval between them (Simon, 1977). Since we know further that the degree of adaptation increases with the intensity of the adaptor (Hillenbrand, 1975; Sawusch, 1977), it is reasonable to suppose that variations in duration, the other determinant of stimulus energy, should have an equivalent effect: the accumulation of anchor/adaptor energy over a test and the resulting degree of effect should be greater for long stimuli than for short. If this is so, we can explain the differences between pitch, loudness and vowels, on the one hand, and stop consonants, on the other, by appealing to the relatively brief duration of the acoustic events that carried the consonantal information and the relatively long duration of the events that carried the pitch, loudness and vowel information.

A virtue of this account is that it meshes neatly with acoustic memory explanations of consonant-vowel differences in other experimental paradigms. Stimulus duration has been shown to be a factor in the variance of discrimination (Fujisaki and Kawashima, 1969; Pisoni, 1973), precategorical acoustic storage (Crowder, 1973; Hall and Blumstein, 1977) and ear advantage in dichotic listening (Godfrey, 1974).

CONCLUSIONS

A fully unified account of the processes underlying the effects of anchoring and adaptation on phonetic and nonphonetic continua is hardly

possible. For even if we rule out various psychophysical processes as factors in phonetic tasks, there is no reason why they should not contribute to effects on nonphonetic tasks. There seem, in fact, to be several processes, any or all of which may contribute in different degrees, depending on the paradigm and the type of task or continuum. As far as the two paradigms are concerned, they do not differ in principle, and, in practice, they are equally "selective." They do, however, differ in anchor/adaptor energy concentration, making "fatigue" more likely to occur under adaptation than under anchoring. Otherwise, there seem to be no grounds for distinguishing the paradigms in terms of the psychophysical processes that they activate.

The cardinal difference between speech and nonspeech continua is that the speech continuum has a fixed perceptual structure: it comes to the listener (or perhaps the listener to it) with its endpoints already anchored. Thus, only the ambiguous center of a speech continuum is liable to the contextual influences that may shape an entire nonspeech continuum. This means that speech is protected from those modes of contextual influence by which an arbitrary parameter of the stimulus distribution, such as its mean or form, comes to serve as the reference for judgments. In short, speech is not subject to the contextual processes modeled by adaptation-level theory or range-frequency analysis.

The two portions of the speech continuum--its rigid ends and loose center--are subject to different main influences: fatigue and contrast. Since fatigue, or desensitization, is not peculiar to speech (see Elliot and Fraser, 1970, for a review), and since the auditory dimensions and values subject to fatigue may be no fewer than all discriminable properties of all sounds, it is misleading to characterize the fatigued entities as "feature detectors." A more neutral, descriptive term, such as "channels of analysis," is to be preferred (Kay and Matthews, 1972; compare Eimas and Miller, in press), particularly if the fatigue is indeed peripheral. By thus turning from a structural claim to a functional description, we resist the temptation to press a physiological metaphor in the absence of converging evidence. The two processes, peripheral fatigue and central contrast, may then be seen as contributing to adaptation effects in nonspeech no less than in speech.

REFERENCES

Ades, A. E. (1974) How phonetic is selective adaptation? Experiments on syllable position and vowel environment. Perception & Psychophysics 16, 61-66.

Ades, A. E. (1976) Adapting the property detectors for speech perception. In New Approaches to Language Mechanisms, ed. by R. J. Wales and E. Walker. (Amsterdam: North Holland).

Ades, A. E. (1977) Source assignment and feature extraction in speech. Journal of Experimental Psychology: Human Perception and Performance 3, 673-685.

Ainsworth, W. A. (1977) Mechanisms of selective feature adaptation. Perception & Psychophysics 21, 365-370.

Bailey, P. (1973) Perceptual adaptation for acoustical features in speech. Speech Perception, Series 2. (Belfast: Department of Psychology, The Queen's University), 29-34.

Bailey, P. (1975) Perceptual adaptation in speech: Some properties of

detectors for acoustical cues to phonetic distinctions. Unpublished Ph.D. dissertation. (Cambridge: University of Cambridge).

Blumstein, S. E., K. N. Stevens and G. N. Nigro. (1977) Property detectors for bursts and transitions in speech perception. Journal of the Acoustical Society of America 61, 1301-1313.

Brady, S. A. and C. J. Darwin. (in press) A range effect in the perception of voicing. Journal of the Acoustical Society of America.

Cole, R. A., W. E. Cooper, J. Singer and F. Allard. (1975) Selective adaptation of English consonants using real speech. Perception & Psychophysics 18, 227-244.

Cole, R. A. and W. E. Cooper. (1977) Properties of friction analyzers for [ʒ]. Journal of the Acoustical Society of America 62, 177-182.

Cooper, F. S., P. C. Delattre, A. M. Liberman, J. M. Borst and L. J. Gerstman. (1952) Some experiments on the perception of synthetic speech sounds. Journal of the Acoustical Society of America 24, 597-606.

Cooper, W. E. (1974) Contingent feature analysis in speech perception. Perception & Psychophysics 16, 201-204.

Cooper, W. E. (1975) Selective adaptation to speech. In Cognitive Theory, vol. 1, ed. by F. Restle, R. M. Shiffrin, N. J. Castellan, H. Lindman and D. B. Pisoni. (Hillsdale, New Jersey: Lawrence Erlbaum Associates).

Cooper, W. E., R. R. Ebert and R. A. Cole. (1976) Perceptual analysis of stop consonants and glides. Journal of Experimental Psychology: Human Perception and Performance 2, 92-104.

Crowder, R. G. (1973) Precategorical acoustic storage for vowels of short and long duration. Perception & Psychophysics 13, 502-506.

Diehl, R. (1975) The effect of selective adaptation on the identification of speech sounds. Perception & Psychophysics 17, 48-52.

Donald, S. L. (1976) The effects of selective adaptation on voicing in Thai and English. Haskins Laboratories Status Report on Speech Research 47, 129-135.

Dorman, M. F., M. Studdert-Kennedy and L. J. Raphael. (1977) Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. Perception & Psychophysics 22, 109-122.

Durlach, N. I. and L. D. Braida. (1969) Intensity perception, I. Preliminary theory of intensity resolution. Journal of the Acoustical Society of America 46, 372-383.

Eimas, P. D. (1963) The relation between identification and discrimination along speech and non-speech continua. Language and Speech 6, 206-217.

Eimas, P. D. and J. D. Corbit. (1973) Selective adaptation of linguistic feature detectors. Cognitve Psychology 4, 99-109.

Eimas, P. D. and J. L. Miller. (in press) Effects of selective adaptation on the perception of speech and visual patterns: Evidence for feature detectors. In Perception and Experience, ed. by R. D. Walk and H. L. Pick, Jr. (New York: Plenum).

Elliot, D. N. and W. R. Fraser. (1970) Fatigue and adaptation. In Foundations of Modern Auditory Theory, vol. 1, ed. by J. V. Tobias. (New York: Academic Press), 117-155.

Foreit, K. G. (1977) Linguistic relativism and selective adaptation for speech: A comparative study of English and Thai. Perception & Psychophysics 21, 347-351.

Fry, D. B., A. S. Abramson, P. D. Eimas and A. M. Liberman. (1962) The identification and discrimination of synthetic vowels. Language and Speech 5,

171-189.

Fujisaki, H. and T. Kawashima. (1969) On the modes and mechanisms of speech perception. Annual Report of the Engineering Research Institute 28, (Tokyo: University of Tokyo), 67-73.

Godfrey, J. J. (1974) Perceptual difficulty and the right-ear advantage for vowels. Brain and Language 4, 323-336.

Goldstein, L. M. and J. R. Lackner. (1973) Alterations of the phonetic coding of speech sounds during repetition. Cognition 2, 279-297.

Hall, L. L. and S. E. Blumstein. (1977) The effect of vowel similarity and syllable length on acoustic memory. Perception & Psychophysics 22, 95-99.

Helson, H. (1964) Adaptation-Level Theory: An Experimental and Systematic Approach to Behavior. (New York: Harper).

Helson, H. (1971) Adaptation-level theory: 1970 and after. In Adaptation Level Theory, ed. by M. H. Appley. (New York: Academic Press), 5-17.

Helson, H. and A. Kozaki. (1968) Anchor effects using numerical estimates of simple dot patterns. Perception & Psychophysics 4, 163-164.

Hillenbrand, J. M. (1975) Intensity and repetition effects on selective adaptation to speech. Research on Speech Perception 2, (Bloomington: Indiana University), 56-137.

Kay, R. H. and D. R. Matthews. (1972) On the existence in human auditory pathways of channels selectively tuned to the modulation present in frequency-modulated tones. Journal of Physiology 225, 657-677.

Lackner, J. R. and L. M. Goldstein. (1975) The psychological representation of speech sounds. Quarterly Journal of Experimental Psychology 27, 173-185.

Liberman, A. M., F. S. Cooper, D. P. Shankweiler and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychological Review 74, 431-461.

Liberman, A. M., K. S. Harris, J. Kinney and H. Lane. (1961) The discrimination of relative onset time of the components of certain speech and nonspeech patterns. Journal of Experimental Psychology 61, 379-388.

Lisker, L. L. and A. S. Abramson. (1964) A cross-language study of voicing in initial stops: acoustical measurements. Word 20, 384-422.

Miller, J. L. (1975) Properties of feature detectors for speech: Evidence from the effects of selective adaptation on dichotic listening. Perception & Psychophysics 18, 389-397.

Miller, J. L. (1977) Properties of feature detectors for VOT: The voiceless channel of analysis. Journal of the Acoustical Society of America 62, 641-648.

Miller, J. L., P. D. Eimas and J. Root. (1977) Properties of feature detectors for place of articulation. Journal of the Acoustical Society of America 61, S48 (A).

Miller, J. L. and P. D. Eimas. (1976) Studies on the selective tuning of feature detectors for speech. Journal of Phonetics 4, 119-127.

Parducci, A. (1965) Category judgement: A range-frequency model. Psychological Review 72, 407-418.

Parducci, A. (1974) Contextual effects: A range-frequency analysis. In Handbook of Perception 2, ed. by E. C. Carterette and M. P. Friedman. (New York: Academic Press), 127-141.

Pisoni, D. B. (1973) Auditory and phonetic memory codes in the discrimination of consonants and vowels. Perception & Psychophysics 13, 253-260.

Pisoni, D. B., J. R. Sawusch and F. T. Adams. (1975) Simple and contingent adaptation effects in speech perception. Research on Speech Perception

2, (Bloomington: Department of Psychology, Indiana University), 22-55.
Repp, B. H. (1977) Dichotic competition of speech sounds: The role of acoustic stimulus structure. Journal of Experimental Psychology: Human Perception & Performance 3, 37-50.
Repp, B. H., A. M. Liberman, T. Eccardt and D. Pesetsky. (1978) Perceptual integration of acoustic cues for stop, fricative and affricate manner. Haskins Laboratories Status Report on Speech Research SR-53, vol. 2, 61-83.
Sarris, V. (1967) Adaptation-level theory: Two critical experiments on the Helson's weighted average model. American Journal of Psychology 80, 331-355.
Sawusch, J. R. (1976) Selective adaptation effects on endpoint stimuli in a speech series. Perception & Psychophysics 20, 61-65.
Sawusch, J. R. (1977) Peripheral and central processing in speech perception. Journal of the Acoustical Society of America 62, 738-750.
Sawusch, J. R. and D. B. Pisoni. (1973) Category boundaries for speech and nonspeech sounds. Journal of the Acoustical Society of America 54, 76 (A).
Sawusch, J. R. and D. B. Pisoni. (1976) Response organization and selective adaptation to speech sounds. Perception & Psychophysics 20, 413-418.
Sawusch, J. R., D. B. Pisoni and J. E. Cutting. (1974) Category boundaries for linguistic and non-linguistic dimensions of the same stimuli. Journal of the Acoustical Society of America 55, S55 (A).
Shankweiler, D. P., W. Strange and R. Verbrugge. (1976) Speech and the problem of perceptual constancy. In Perceiving, Acting and Comprehending: Toward an Ecological Psychology, ed. by R. Shaw and J. Bransford. (Potomac, Maryland: Erlbaum).
Sherif, M., D. Taub and C. I. Hovland. (1958) Assimilation and contrast effects of anchoring stimuli on judgments. Journal of Experimental Psychology 55, 150-156.
Simon, H. J. (1977) Anchoring and selective adaptation of phonetic and nonphonetic categories in speech perception. Unpublished Ph.D. dissertation. (New York: City University of New York).
Stevens, K. N., A. M. Liberman, M. Studdert-Kennedy and S. E. G. Öhman. (1969) Cross-language study of vowel perception. Language and Speech 12, 1-23.
Strange, W. and J. J. Jenkins. (in press) The role of linguistic experience in the perception of speech. In Perception and Experience, ed. by R. D. Walk and H. L. Pick, Jr. (New York: Plenum).
Studdert-Kennedy, M. (1976a) Speech Perception. In Contemporary Issues in Experimental Phonetics, ed. by N. J. Lass. (New York: Academic Press), 243-293.
Studdert-Kennedy, M. (1976b) Stimulus range as a determinant of phoneme boundaries along synthetic consonant continua. Journal of the Acoustical Society of America 60, S92 (A).
Volkmann, J. (1951) Scales of judgment and their implications for social psychology. In Social Psychology at the Crossroads, ed. by J. H. Rohrer and M. Sherif. (New York: Harper).
Warren, R. M. and R. L. Gregory. (1958) An auditory analogue of the visual reversible figure. American Journal of Psychology 71, 612-613.
Woodworth, R. S. (1938) Experimental Psychology. (New York: Holt).

Speech Across a Linguistic Boundary: Category Naming and Phonetic Description*

Leigh Lisker+

ABSTRACT

Crosslanguage testing of speech materials provides a method of checking on hypotheses concerning the properties said to characterize the phonetic elements of one or both of the languages involved in the comparison--the language of the speaker providing the test stimuli and the language of the listener asked to give labeling responses to them. The American English initial stops /ptk/ are described as voiceless fortis aspirated. To help decide whether all these properties must be present if a phonetically naive speaker of English is to label a stop as a member of the /ptk/ set, a group of such persons was asked to identify monosyllables produced by a Dutch speaker as /ptk/ or /bdg/. The Dutch /ptk/ stops, which are voiceless fortis inaspirates, were identified overwhelmingly as /ptk/; from this it appears that aspiration may not be a necessary requirement for /ptk/ judgments from English speakers, even though it regularly occurs in English initial /ptk/. A similar test involving Korean monosyllables revealed that stops described as voiceless lenis aspirates (moderately so) also elicited /ptk/ responses. Thus one might infer that fortis articulation or aspiration is sufficient for English /ptk/, and that the English set is significantly less fortis than Dutch /ptk/, and not significantly more fortis than Korean /p′t′k′/.

INTRODUCTION

It is commonly believed that almost any vocal tract, no matter what the ethnic affiliation of its owner, is inherently able to function "natively" in any language community, so long as that tract, and the ear to which it is attached, are "normal" and have been welcomed into that community at a "normal" age, namely in infancy. Linguistic inabilities, including phonetic,

---

*A slightly different version of this paper was presented before the IPS-77, the International Phonetics Sciences Congress, December 1977, Miami Beach, Florida.

+Also University of Pennsylvania.

Acknowledgment: This research was supported in part by the National Institute of Neurological and Communicative Disorders and Stroke, Grant NS-13870.

[HASKINS LABORATORIES: Status Report on Speech Research SR-54 (1978)]

that are manifested in later life, are less evenly distributed over individuals, but presumably are in part culturally determined; some Americans, for example, speak more acceptable (to the French) French than others, but there is a recognized American-accented French. The nature of these phonetic inabilities is not all that well understood, for we are still not clear about what is perceptually based and what is a matter of more or less arbitrary category naming. Once acoustic signals are apprehended as speech, their attributes seem to be evaluated by reference to a vocal tract that might have produced them, and beyond that, they are labeled in terms of categories given by the language in which that vocal tract is speaking, provided the listener shares the language of the speaker. For the naive listener, who by definition has available to him only the categories of his own language, the categories of the speaker's language, if it happens to be different, should have no bearing on his labeling behavior. Comparing the native and nonnative labelings of speech samples, however, enables us to map the categories of one language on another, and also serves as some check on hypotheses regarding the phonetic basis for category distinctions in one or both of the languages being compared.

STOP CATEGORIES AND ARTICULATORY FORCE

Let us consider the crosslanguage correspondences of some stop consonant categories. English stops in initial position have been characterized differentially with respect to the phonetic features of voicing, aspiration and level of articulatory force. The measure of voice onset timing (VOT) has provided data to suggest that the /bdg/ and /ptk/ categories differ significantly, in the statistical sense, in their VOT values. In addition, experiments in synthesis and the systematic manipulation of normally produced speech signals have yielded no strong evidence to discount the perceptual importance of this VOT dimension. Since the measure relates to the features of both voicing and aspiration, this leaves the force-of-articulation features out in the cold. The relation between a postulated dimension of articulatory force and other features recognized by the phonetician is a somewhat obscure one, for it is not the case that force of articulation is simply another phonetic dimension, like voice or tongue height, for example. Rather, it is a feature that is brought into phonetic description in order to explain how some of these other more readily observed and measured properties are generated, particularly where they occur as properties of phonologically identical but phonetically different events. Thus the partially alternating properties of aspiration and relatively longer closure duration of English /ptk/ have been referred to a "fortis" level of articulatory force, while the contrasting categories are "lenis," a designation that is said to explain why initial voiceless unaspirated and medial voiced stops are grouped together in the /bdg/ set. In very much the same way, in Korean, lenis articulation has been asserted (Kim, 1965) to be the property underlying a phonological class that includes voiceless stops with a moderate degree of aspiration (or perhaps murmur, if we follow Ladefoged, 1971) as well as quite ordinary voiced stops.

In some languages it seems that voiced and voiceless stops are, ipso facto, lenis and fortis respectively. However, there have been cited (Ladefoged, 1971; Catford, 1977) languages in which the dimension of articulatory force is said to operate quite independently of any voicing difference. The argument for (or against) an independent fortis-lenis dimension is complicated

by the fact that some writers on the subject have shown little tendency to restrict their choice of physical indices of articulatory force to properties that are clearly independent of voicing. Of course the terms "fortis" and "lenis" have a useful function, in that, as qualifiers not well enough defined to be demonstrably inapplicable to the stops of a specific language such as English, they can serve: 1) as category names acceptable to those who are unconvinced that only a voicing contrast is present, and 2) as the cover term for any observable features other than voicing that show significant differences between distinct categories. Those already convinced take a demonstration that any such difference exists as proof of the fortis-lenis nature of the contrast. One investigator who has written extensively on the subject has, after a long hunt for indices that would yield the "right" answer, finessed the question by supposing that the incontrovertible evidence for a fortis-lenis difference is the fact that phonetically naive subjects regularly report /ptk/ to be harder to produce that /bdg/, and that <u>this</u> difference rests on a proprioceptive sensitivity to the greater intraoral air pressures developed during /ptk/ (Malécot, 1970).

Despite all the doubt expressed about a dimension of articulatory force as a phonological feature of specific languages, it seems to be obviously true that a speaker, say of English, is capable of regulating the degree of force with which the lips come together during a /p/ or /b/ (or /m/) occlusion, and the stops differing in this feature can properly be said to differ in force of articulation. Moreover, it does not appear unreasonable to suppose that, despite intra- and interspeaker variation for a single language, there may be differences between languages in the mean mechanical pressures exerted during the production of such stop consonants. Thus, for example, the initial voiceless stops of Dutch, which are unaspirated in the standard dialect, appear to be produced with a good deal of energy; in my judgment they can be plausibly labeled [+fortis] as compared with the Dutch voiced stops, or for that matter, as compared with the voiceless aspirates of American English. The initial voiceless inaspirates of Korean, which Kim (1965) asserts to be <u>phonologically</u> [+tense] (the same thing as [+fortis]), also seem to be produced with a good deal of energy, though perhaps less than is involved in producing the phonetically comparable Dutch stops.

The situation in English is more complex than I earlier suggested. For one thing, the famous case of post-/s/ stops is not entirely clear--they are traditionally considered to be varieties of /ptk/: voiceless, unaspirated, of uncertain degree of force, though perhaps fortis. <u>If</u> they are fortis, then this attribute is not sufficient to result in /ptk/-labelings by English-speaking listeners when the /s/-noise is stripped away by tape-editing (Lotz, Abramson, Gerstman, Ingemann and Nemser, 1960). If /ptk/ are distinctively [+fortis], and if the post-/s/ stops are /ptk/, then removal of the /s/-noise should yield /ptk/ rather than /bdg/. If it is argued that the post-/s/ stops are neutral as to force of articulation, since there is only a single set of stops--one for each place of articulation--then there is still the problem of medial /ptk/ before unstressed vowels. These stops are also reported as /bdg/ when editing puts their releases in initial position. A survey of the phonetic literature on English indicates that there is not complete agreement as to whether the /p/ of <u>rapid</u>, for example, is fortis or lenis. If it is considered to be fortis, while /b/ is lenis, this fortis quality does not prevent listeners from identifying it as /b/ following removal of the

ENGLISH

DUTCH	/ B	D	G	P	T	K /
/ B	99			1		
D		100				
G			100			
P	13			86	1	
T		9		2	89	
K /			3			96

Figure 1: Responses of eight English-speaking subjects to ten tokens of Dutch /ba da ga pa ta ka/; two responses per subject per token; subjects asked to label with English category names; percentage responses.

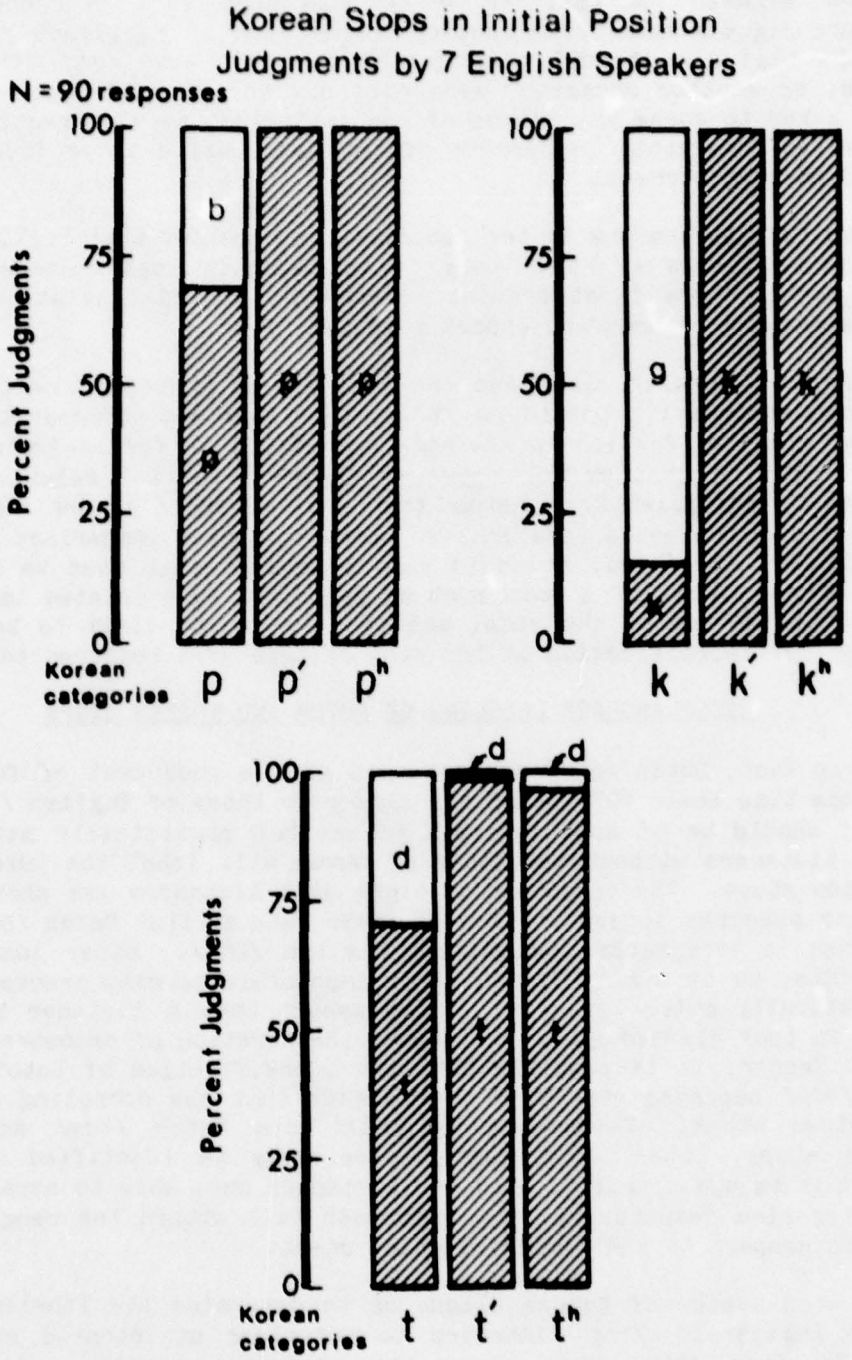


Figure 2: Assignment by English-speaking listeners of Korean stop categories (three tokens of each) /ptk p' t' k' $p^h t^h k^h$/ to the English categories /bdg ptk/.

preclosure signal. A test in which listeners were presented with the postclosure signals from three recoded tokens each of rapid and rabid yielded the result that all stimuli were judged to begin with /b/. Moreover, when listeners, on another occasion, were told how the stimuli had been prepared and were asked to guess the source of each stimulus, they correctly identified those derived from rabid 70 percent of the time, while those from rapid were judged 43 percent correct.

These results conform to the generally held belief that English listeners accept initial stops as /ptk/ only if voice onset lags release by some 35 msecs or more. There is at present no commonly shared conviction as to what listeners require in order to report a medial /ptk/.

If Engish post-/s/ stops and the postrelease phases of medial voiceless unaspirated stops are reported as /bdg/, this does not necessarily invalidate the belief that the English /ptk/-/bdg/ opposition is fortis-lenis in nature. Thus it might be that medial /ptk/, although [+fortis] relative to medial /bdg/, is not sufficiently stronger than initial /bdg/ to be separated from the latter when presented in a context allowing direct comparison with initial stops. On the other hand, it could also be argued that once we have removed the preclosure signal of a word such as rapid, we have deleted important cues to the fortis nature of the stop, and that we cannot claim to be presenting medial /p/ for identification in the kind of test just referred to.

CROSSLANGUAGE LABELING OF DUTCH AND KOREAN STOPS

If, in fact, Dutch /ptk/ are produced with a good deal of force, and if at the same time their VOT values are closer to those of English /bdg/ than of /ptk/, it should be of some interest to see how phonetically naive English-speaking listeners without knowledge of Dutch will label the Dutch voiceless unaspirated stops. The responses of eight such listeners are shown in Figure 1, and one possible interpretation of these data is that Dutch /ptk/ are more fortis than is acceptable for initial English /bdg/. Other interpretations are possible, to be sure. First, it is impossible to make precise the notion of "phonetically naive listener," or to assume that a listener so described remained in that blessed state throughout the duration of exposure to the test stimuli. Second, it is possible that the identification of Dutch /ptk/ with English /bdg/ depended crucially on the fact that the competing stimuli were fully voiced stops. In competition with both Dutch /bdg/ and voiceless aspirated stops, Dutch /ptk/ might conceivably be identified with English /bdg/. What is undeniable is that our listeners were able to separate the two Dutch categories despite the fact that both fall within the range of English /bdg/ with respect to the timing of voice onset.

The stop system of Korean allows us to determine the labeling responses of naive English-speaking listeners to voiceless unaspirated stops (called "tense" by Kim, 1965) when these are presented together with voiceless aspirates. In addition, we can discover whether the so-called lenis voiceless stops will be classed with English /bdg/ or /ptk/; if the former, we may suppose it is on the basis of a shared "lenishness," if the latter, it is because of the similarity in VOT values. From the responses shown in Figure 2, it appears that Korean /p/ and /t/ are assigned largely to English /pt/, despite the inclusion of voiceless aspirated stops in the same test. Unlike

the Dutch case, about 30 percent of the responses were /bd/, a fact that might attribute either to the presence of the aspirates, or to a possible difference in the force with which the Korean and Dutch voiceless inaspirates are articulated. Korean /k/ is very differently labeled, although there is no reason to think that it is less strongly articulated than /pt/. If Korean /ptk/ are all articulated so as to produce strong release bursts, then possibly the readiness to accept Korean /k/ as English /g/ is explained by the fact that English /g/, with its relatively long delay in voicing onset, has a stronger burst than English /bd/.

The so-called middle category of Korean stops, the "lenis" somewhat aspirated voiceless stops found in initial position, are assigned entirely to English /ptk/. They are either not lenis enough to satisfy the requirements for English /bdg/ (although the "fortis" Korean /ptk/ did elicit a significant number of /bd/ and especially /g/ responses), or perhaps English /ptk/ are not especially fortis, at least when there is some aspiration (even if it is "murmur").

CONCLUSION

In summary, the labelings of English speakers asked to assign English stop category names to Dutch and Korean initial stops indicate that the voiceless unaspirated, and possibly fortis, stops of the two latter languages are not categorized on the basis of their VOT values, at least as these are determined by acoustic measurement. If the features determining their classification are not of laryngeal origin, then we may suppose that other acoustic features, which might be associated with a high level of articulatory force, are responsible for the observed behavior. The evaluation of Korean /p't'k'/, on the other hand, suggests that a high level of force is not a prerequisite for English /ptk/. Thus it appears that, assuming we accept the validity of assertions regarding the fortis-lenis character of the foreign stop categories dealt with, English initial /ptk/ may be cued either by aspiration (that is, a lag in voicing onset) or by some other features, yet unspecified, produced by fortis articulation, while English /bdg/ may require an absence of both aspiration and the acoustic consequences of fortis production. It is not entirely impossible that the features that led our listeners to associate the Dutch and Korean voiceless inaspirates with English /ptk/ are dependent upon the nature and timing of laryngeal adjustments during the stop articulations.

REFERENCES

Catford, J. C. (1977) Fundamental Problems in Phonetics. (Edinburgh: Edinburgh University Press).

Kim, C.-W. (1965) On the autonomy of the tensity feature in stop classification. Word 21, 339-359.

Ladefoged, P. (1971) Preliminaries to Linguistic Phonetics. (Chicago: University of Chicago Press).

Lotz, J., A. S. Abramson, L. J. Gerstman, F. Ingemann and W. S. Nemser. (1960) The perception of English stops by speakers of English, Spanish, Hungarian and Thai: A tape-cutting experiment. Language and Speech 3, 71-77.

Malécot, A. (1970) The lenis-fortis opposition: Its physiological parameters. Journal of the Acoustical Society of America 47, 1588-1592.

Discrimination of Subphonemic Phonetic Distinctions

S. Lea Donald+

ABSTRACT

The purpose of the experiments reported here was to determine whether discrimination depends on actual phonemic categorization or whether it can rely on an awareness of a phonetic distinction that is used phonemically in a subject's language, but which is not distinctive in the context being tested. Seven native speakers of Thai and seven native speakers of English took part in a set of discrimination experiments. The Thai-speaking subjects took part in discrimination tests of both labial and velar stimuli that varied along voice onset time (VOT) continua. The English-speaking subjects took part only in a velar discrimination test. The Thai language makes phonemic distinctions between voiced and voiceless unaspirated stops at the labial and dental places of articulation. However, Thai does not make this distinction at the velar place of articulation. English does not make this utterance-initial distinction at any place of articulation. Therefore, the Thai velar discrimination functions can be compared both with discrimination functions in which a phonemic distinction is made between voiced and voiceless unaspirated stimuli and also with the English speakers' discrimination functions where no such phonemic distinction exists in the language as a whole.

INTRODUCTION

The purpose of the experiments reported here was to determine whether discrimination depends on actual phonemic categorization or whether it can rely on an awareness of a phonetic distinction that is not phonologically distinctive in the context being tested.

Streeter (1976) conducted an experiment along these lines. She tested the ability of native speakers of Kikuyu to discriminate among labial and apical stops along a voice onset time (VOT) continuum. Streeter states that Kikuyu makes a phonological distinction between "prevoiced" and "voiced" apical stops [presumably (pre-)voiced and voiceless unaspirated, in normal phonetic terminology]. Kikuyu speakers reliably discriminated between apical stops across a boundary at approximately 10 msec. Since Streeter presents no identification functions to compare with the discrimination functions, it is difficult to judge whether the discrimination peak is mediated by the phonemic identification of the stimuli involved.

+Also University of Connecticut, Storrs.

[HASKINS LABORATORIES: Status Report on Speech Research SR-54 (1978)]

Streeter herself explains these results in terms of subjects' sensitivity to qualitative acoustic changes along the VOT continuum. Specifically, she suggests that subjects may be discriminating on the basis of the presence or absence of a detectable first formant.

Although according to Streeter, Kikuyu has only a prevoiced labial stop and not a voiceless unaspirated labial stop,[1] Kikuyu subjects produced two peaks in their discrimination of labial stops: one peak in the region of -20 to -10 msec, and the second in the region 10 to 25 msec. As for the apical data, Streeter primarily attributes the discrimination peaks obtained here to acoustic discontinuities along the VOT continuum. It is puzzling that Kikuyu subjects' performance on labial discrimination was better than their performance on apical discrimination. Not only is there an additional peak of sensitivity for the labial stops, but the shared peak is greater for the labials than for the apicals. Perhaps the phonemic distinction in the apicals obscures the auditory effect there and thus promotes a purely phonetic effect.

Due to the lack of identification functions with which to compare the obtained discrimination functions, it is unclear how to assess the relative contributions of the phonemic and auditory distinctions involved in Streeter's experiment.

Abramson and Lisker (1968) found that Thai listeners were able to discriminate between voiced and voiceless unaspirated labial and dental stop stimuli. In velars, for which there is no corresponding voiced-voiceless unaspirated distinction, they found some evidence of slightly heightened discrimination around VOT values corresponding to the labial and dental voiced-voiceless unaspirated boundaries; but with data from only two subjects, any conclusions are premature.

EXPERIMENT

The experiment reported here addresses the question of whether Thai-speaking subjects are able to discriminate between velar stimuli across a region corresponding to the labial and dental voiced-voiceless unaspirated boundaries. Thai-speaking subjects and English-speaking subjects took part in a velar discrimination task. Additionally, the Thai subjects took part in a labial discrimination task. Therefore, the Thai velar discrimination functions can be compared both with discrimination functions in which a phonemic distinction is made between voiced and voiceless unaspirated stimuli (the Thai labial condition), and also with discrimination functions where no such utterance-initial phonemic distinction exists in the language as a whole (the English velar condition).

---

[1]The phonological structure of Kikuyu may not be as simple as Streeter implies. Jones (1950) states that [b] is not used except after [m] [for example, in [mbori] (goat)], in at least one dialect of Kikuyu. Unfortunately, Streeter does not identify the dialect of her subjects.

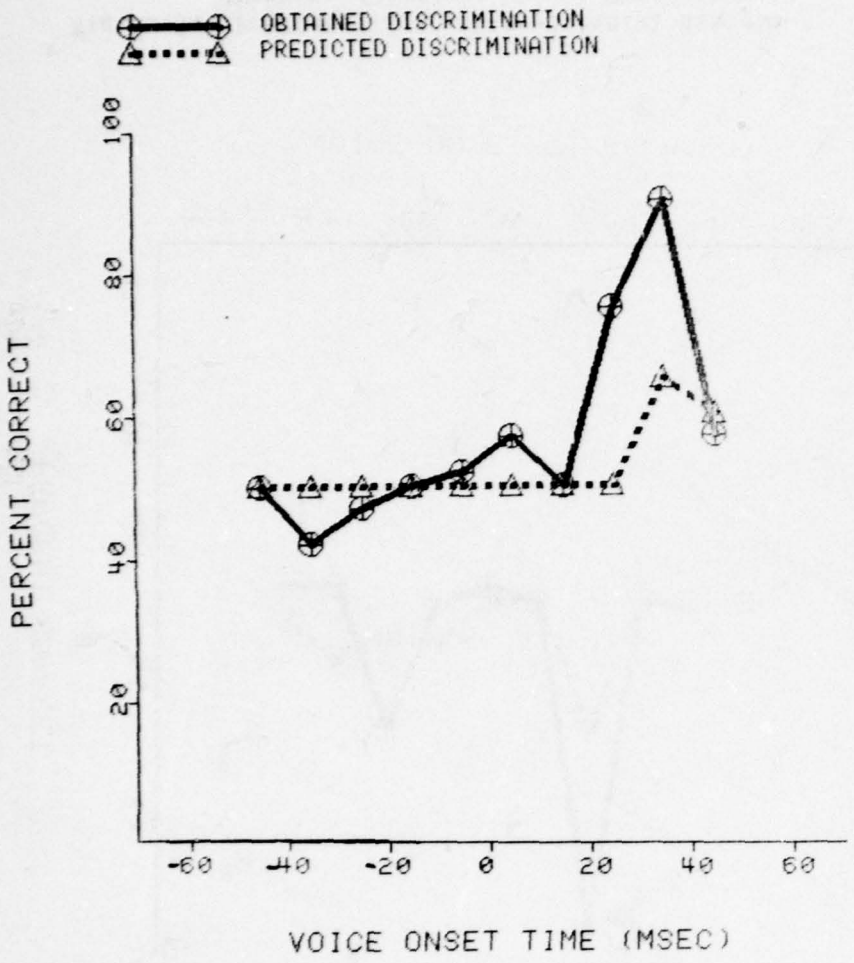


Figure 1: Predicted and obtained velar discrimination functions for an English-speaking subject.

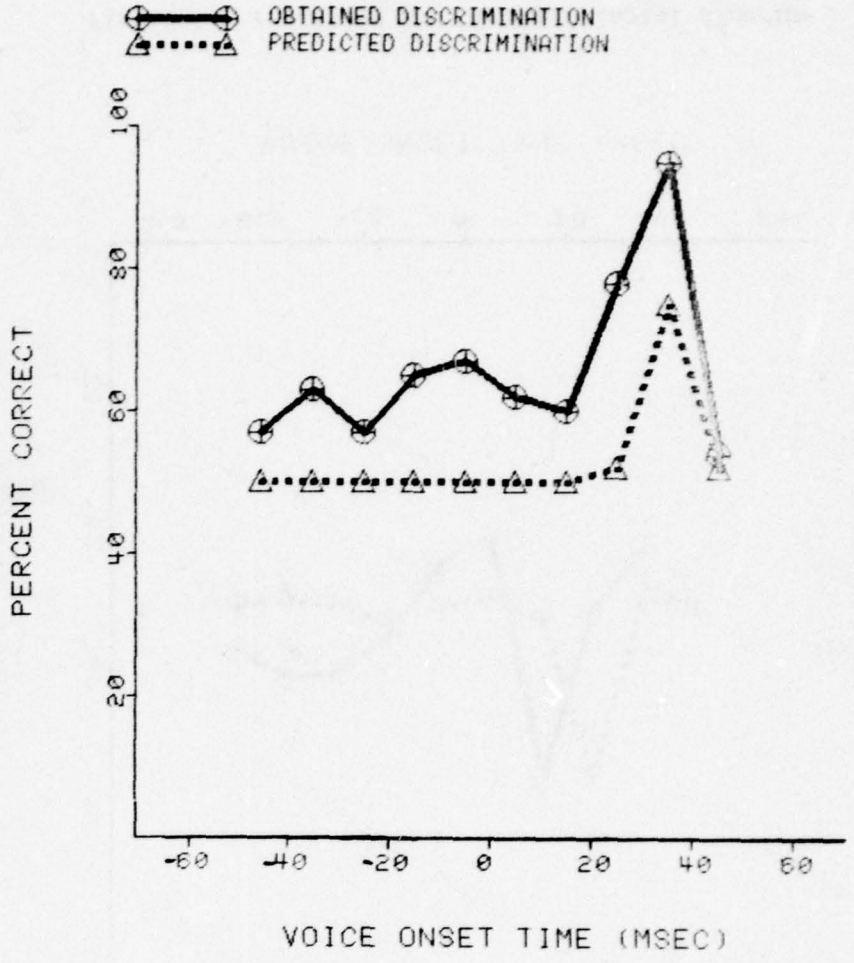


Figure 2: Predicted and obtained velar discrimination functions for an English-speaking subject.

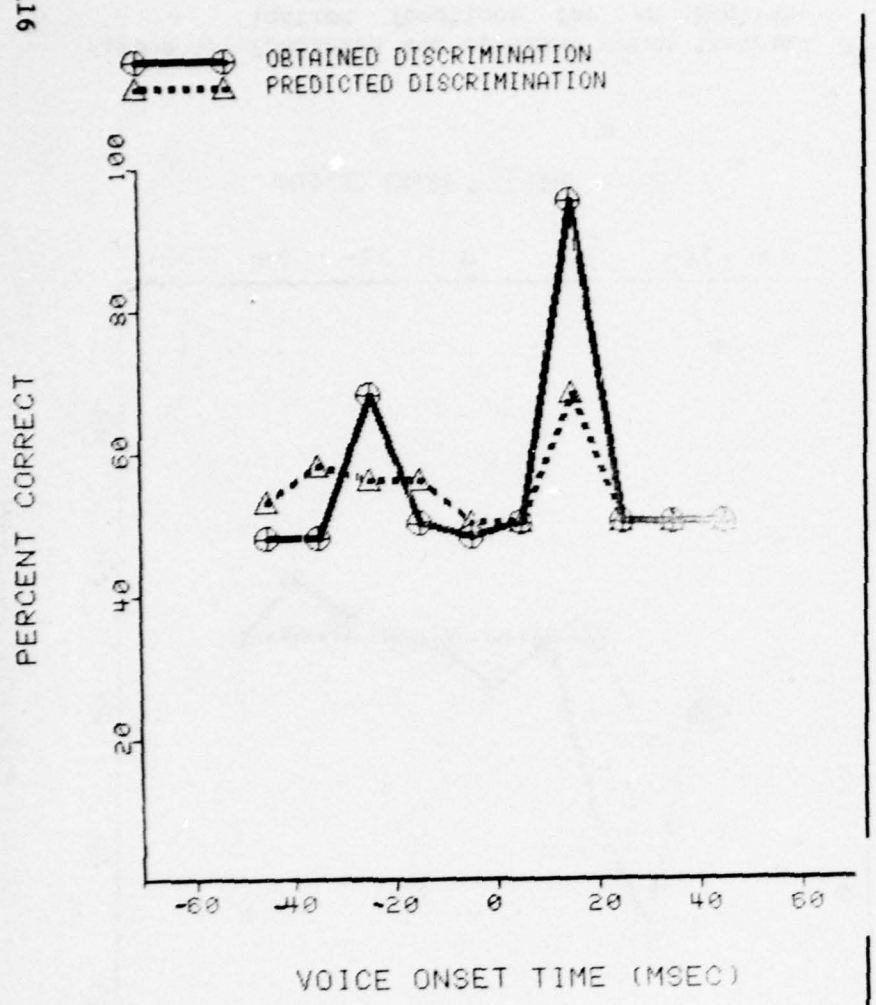


Figure 3: Predicted and obtained labial discrimination functions for a Thai-speaking subject.

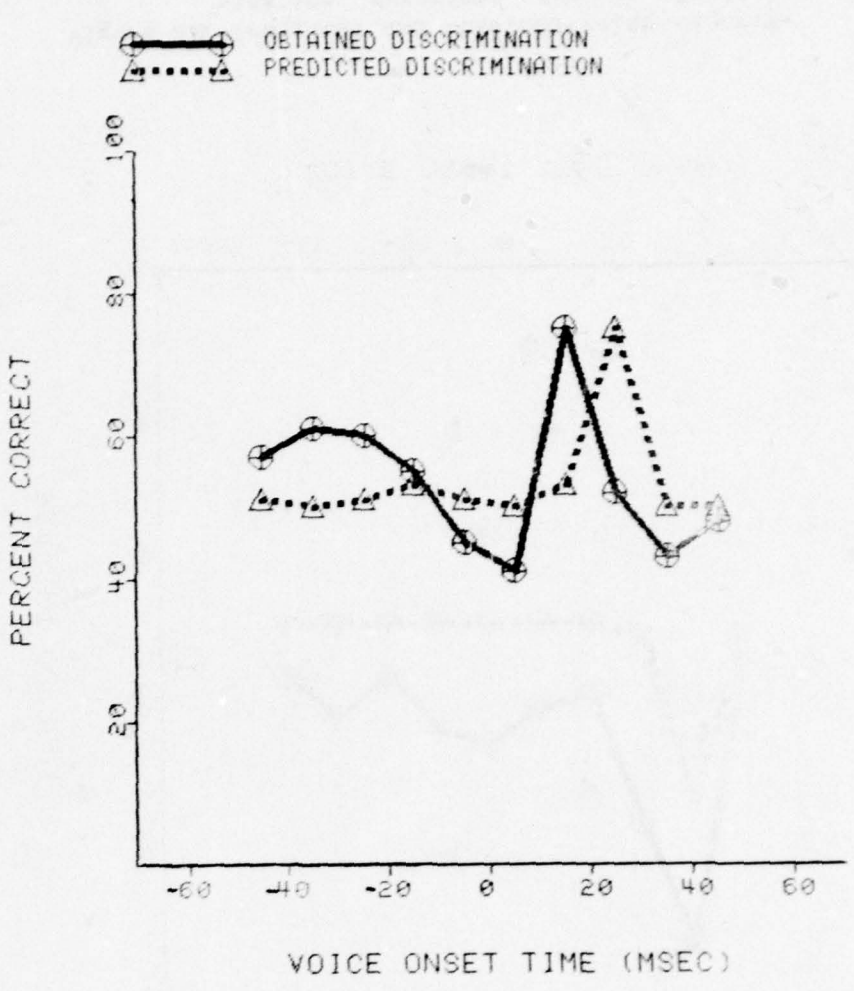


Figure 4: Predicted and obtained labial discrimination functions for a Thai-speaking subject.

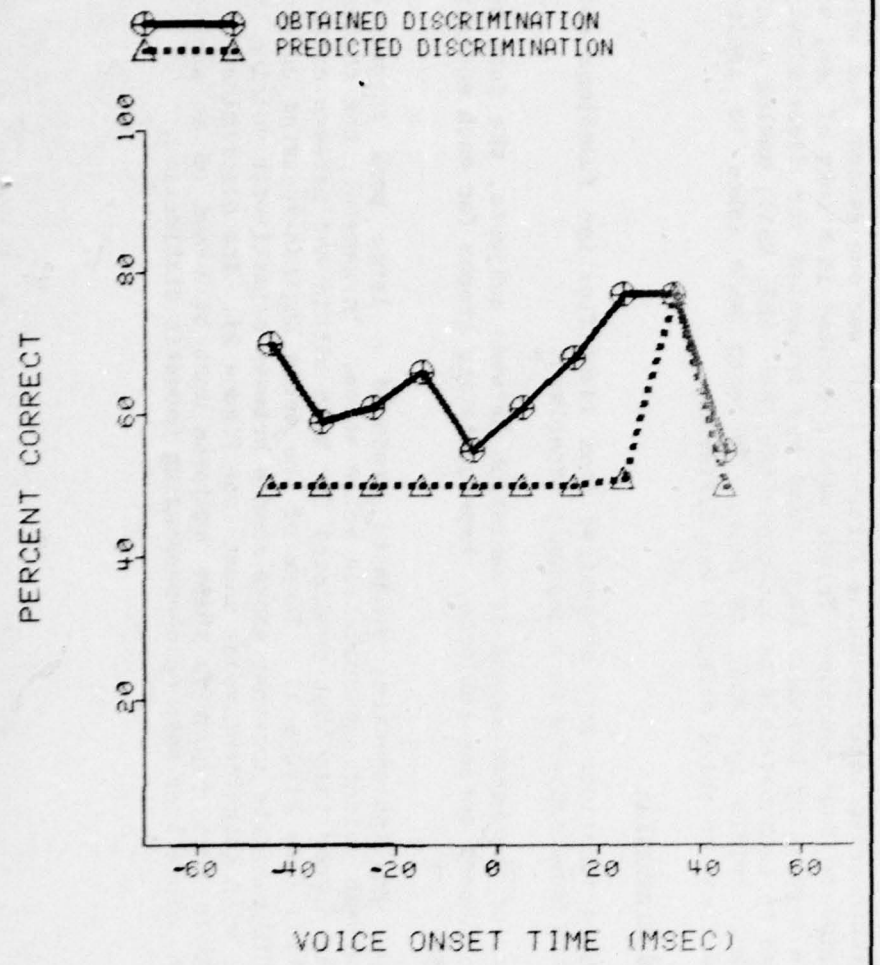


Figure 5: Predicted and obtained velar discrimination functions for a Thai-speaking subject.

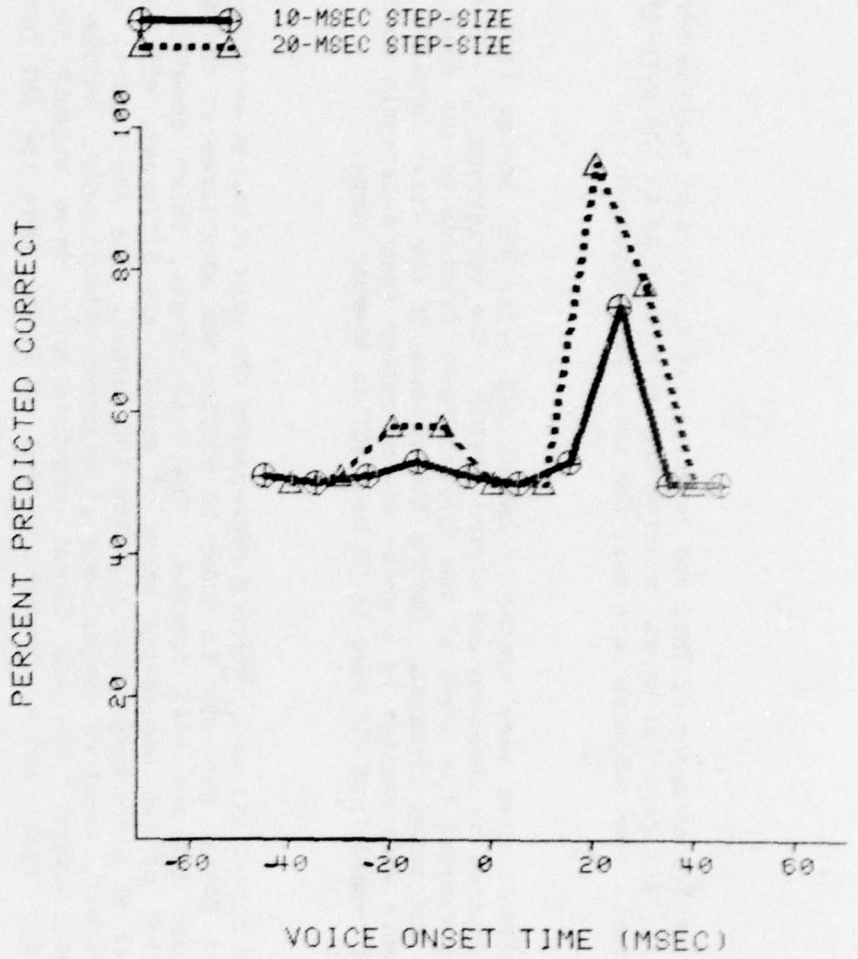


Figure 6: Labial discrimination functions predicted for 10-msec stepsize and 20-msec stepsize for a Thai-speaking subject.

Subjects

Seven native speakers of Thai and seven native speakers of English served as subjects. All reported normal hearing and were naive as to the purpose of the experiment. The subjects were paid for their participation.

Stimuli

The stimuli used were synthetic labial and velar VOT series from the continua prepared by Abramson and Lisker (1970). The variations in VOT were produced by varying the onset of the first formant relative to the onset of the second and third formants. During the absence of the first formant, the upper formants were excited by a noise source rather than a periodic source. The stimuli ranged from -50 msec to 50 msec VOT in 10-msec steps.

Procedure

An AXB format was used, where X represented the same stimulus as either A or B. This format was used in order to combine the advantages of the more commonly used ABX and 4IAX formats. The 4IAX format, which consists of judging which of two successive pairs of stimuli are different, elicits a higher level of performance than does the ABX format. The AXB format, like the 4IAX format, requires comparisons of adjacent stimuli only. Unlike the 4IAX format, however, the AXB format requires only three stimuli to be presented per trial, and so requires a test no longer than the ABX format requires.

The within-triad inter stimulus interval (ISI) was one second, and triads were separated by four seconds. Triads were presented in blocks of ten, with ten seconds separating blocks. Each triad was presented for discrimination eleven times in each possible permutation (AAB, BAA, ABB, BBA), making a total of forty-four trials for each AB pair. Subjects were asked to indicate whether the first or third stimulus was different.

Results and Discussion

Predicted functions were determined from identification functions obtained from the same subjects in a separate experiment.

Because of the large amount of variation between subjects, the data have not been collapsed across subjects. Representative graphs for each condition are included.

All the English-speaking subjects produced a large peak across the boundary between voiced and voiceless velar stops. In general, the obtained function was higher than that predicted for both within and between category discrimination (see Figure 1). Three of the seven English-speaking subjects appear to discriminate somewhat above chance between stimuli with voicing lead and stimuli with coincident voice onset (see Figure 2). The discrimination of VOT stimuli in this region by these subjects must be based on an auditory distinction, since they have no corresponding phonetic distinction.

In the Thai labial discrimination condition, a sizable discrimination peak was predicted and obtained across the boundary between voiceless unaspirated and voiceless aspirated labial stops. For four of the seven subjects, however, the obtained discrimination peak fell 10 msec to the left of that VOT value predicted from the identification data (see Figure 3).

Due to the interaction of the 10 msec comparison size and the slope of the identification functions, the predicted discrimination peaks across the boundary between prevoiced and voiceless unaspirated stops for certain subjects are very small. By the formula used for predicting discrimination, a 40 percent difference in the identification function is necessary to produce a 10 percent rise in the discrimination peak. Six of the seven subjects, however, did produce a distinct discrimination peak across this boundary. The seventh subject showed a slight rise, although the predicted function shows no variation (see Figure 4).

As reported in earlier studies (Abramson and Lisker, 1968, 1970), the obtained discrimination peak across the boundary between voiced and voiceless unaspirated stops is lower than that obtained across the boundary between voiceless inaspirates and voiceless aspirates (for four of the seven subjects). The same generalization, however, holds true for the predicted discrimination peaks: the predicted discrimination peak across the boundary between voiced and voiceless unaspirated stops is lower than that across the boundary between voiceless inaspirates and voiceless aspirates for six of the seven subjects.

There are two possible reasons for this discrepancy. Abramson and Lisker (1968) suggest that this is because aspiration noise differences are better detected than low-frequency, low-intensity voicing lead differences. However, it is also possible that the synthetic stimuli used do not reflect the voiced-voiceless unaspirated distinction as well as they reflect the voiceless unaspirated-voiceless aspirated distinction. This interpretation is supported by the fact that one Thai subject had to be eliminated from the original subject pool because he was not able to identify any of the labial stimuli as voiced stops.

In the Thai velar discrimination condition, all subjects produced a sizable discrimination peak across the voiceless unaspirated-voiceless aspirated boundary. Six of the subjects also showed a second discrimination peak crossing a possible voiced-voiceless unaspirated boundary (see Figure 5). Since Thai has no phonemic distinction between voiceless unaspirated and voiced velars, no peak is predicted across this distinction. For five of these subjects, the peak falls at VOT values of -15 or -25 msec. For all but one of these five subjects, this discrimination peak falls at precisely the same VOT value as the corresponding discrimination peak in the labial data. For the fifth subject, the two peaks are offset by only 10 msec. For the sixth subject, this peak falls at 5 msec. This peak is not inappropriate if this subject were discriminating between the presence and absence of voicing. The discrimination peak for this subject on the labial continuum between voiced and voiceless unaspirated stops fell close to that same value, at -5 msec.

Summary

In summary, this study provides suggestive evidence that Thai subjects have some awareness of the subphonemic distinction between voiced and voiceless unaspirated velar stops. Data collected in a pilot study indicated that a 10-msec comparison size was large enough to elicit readily interpretable data. However, the fact that the predicted discrimination peaks across the boundary between labial voiced and voiceless unaspirated stimuli for the Thai subjects are so low suggests that the 10 msec comparison size may have presented too difficult a task. One subject remarked after taking part in her first identification test of velars, that some of the stimuli presented were like Thai voiced labial and dental stops, but that the Thai language did not use that phonetic value. That this subject did not show a particularly robust peak separating velar stops with voicing lead from stops with coincident voice onset clearly indicates that the task was too difficult to pick up a distinction of which she was explicitly aware.

Previous studies (Abramson and Lisker, 1968, 1970; Streeter, 1976; Williams, 1977) have not used a 10-msec step size but instead have used 20 msec, 30 msec, and even 40 msec comparisons. Figure 6 compares the discrimination functions for a single subject predicted for a 10-msec step size and for a 20-msec step size. This figure clearly illustrates the greater difficulty presented by a 10-msec comparison size over a 20-msec comparison size. Considering the difficulty of the task, it is perhaps more surprising that subjects performed as well as they did than that performance was not higher. Use of a larger step size would presumably produce larger discrimination peaks, and thus demonstrate more clearly that Thai subjects are aware of the subphonemic distinction between prevoiced and voiceless inaspirate velar stops.

REFERENCES

Abramson, A. S. and L. Lisker. (1968) Voice onset timing: Cross language experiments in identification and discrimination. Haskins Laboratories Status Report on Speech Research SR 13/14, 49-63.

Abramson, A. S. and L. Lisker. (1970) Discriminability along the voicing continuum: cross-language tests. Proceedings of the 6th International Congress of Phonetic Sciences, Prague, 1967. (New York: Academia), 569-573.

Jones, D. (1950) The Phoneme: Its Nature and Use. (Cambridge: W. Heffer and Sons, Ltd.).

Streeter, L. A. (1976) Kikuyu labial and apical stop discrimination. Journal of Phonetics 4, 43-49.

Williams, L. (1977) The voicing contrast in Spanish. Journal of Phonetics 5, 169-184.

Anticipatory Coarticulation: Some Implications from a Study of Lip Rounding*

Fredericka Bell-Berti+ and Katherine S. Harris++

ABSTRACT

The anticipation of articulatory features, in particular lip rounding in anticipation of a rounded vowel, has been reported to occur as many as four segments before the segment for which the feature is specified. In the data presented here, we find that the motor commands for the rounding gesture for /u/ begin a fixed time before the onset of the vowel. This timing is unaffected by the number of preceding consonant segments or the location of syllable boundaries in the preceding string. Thus, the initiation of lip rounding appears to be linked to other features of the vowel articulation.

INTRODUCTION

This paper examines the well-known effect of rounded vowels on preceding unrounded consonants--the "anticipatory coarticulation" of the vowel lip-rounding gesture during the preceding consonant articulation.

Two explanations have been proposed for anticipatory coarticulation. One by Kozhevnikov and Chistovich proposes that speech is organized in articulatory syllables of some kind, so that the syllable boundary provides a limit on anticipatory coarticulation. This limit, that is, the syllable boundary, is described as the beginning of a consonant string preceding a vowel. It has been shown several times that coarticulation can extend across conventionally defined syllable boundaries (for example, Daniloff and Moll, 1968; McLean, 1973; Benguerel and Cowan, 1974).

A second explanation of anticipatory coarticulation has been proposed by Henke (1967), the "look-ahead" model. In this model speech units are organized as bundles of independent parallel articulatory features, planned in parallel, with no restriction on the initiation of a feature of some

---

*Some of these data were presented at the 92nd Meeting of the Acoustical Society of America, San Diego, California, November 1976, and a version of this paper was presented at the 94th Meeting of the Acoustical Society of America, Miami, Florida, December 1977.

+NIH post-doctoral fellow, on leave from Montclair State College, Upper Montclair, New Jersey, September 1977 through August 1979.

++Also The Graduate School, The City University of New York.

[HASKINS LABORATORIES: Status Report on Speech Research SR-54 (1978)]

downstream segment except that it not be antagonistic to some intervening segment; thus, the syllable boundary has no particular status and should not inhibit coarticulation. This view is supported by the evidence, cited above, that coarticulation can occur across syllable boundaries, but is not supported by two electromyographic studies (Bell-Berti and Harris, 1974; Gay, 1975), that show that vowel-related activity is suppressed during the production of an apparently nonantagonistic intervening consonant gesture. Both these models of coarticulation are organized around phonetic units--articulatory syllables in the Kozhevnikov and Chistovich model, and segmental features in the Henke model--and have not considered the temporal course of articulatory events. The purpose of this study is to reexamine the anticipatory coarticulation of lip rounding with a more thorough consideration of the duration of acoustic segments.

METHODS

In the first study, we examined the time course of orbicularis oris activity for the rounded vowel /u/ when it was preceded by /s/, /t/, /st/, and /ts/. These consonants, in turn, were preceded by either /i/, described as having a spread lip position, or by /a/, described as having a neutral lip position.

Two subjects repeated the eight nonsense utterances, which were of the form [pistup] and [patup], 16 to 18 times each. One subject embedded the items in a carrier phrase, "Now say _____ again." EMG activity was recorded from the orbicularis oris, with surface electrodes (Allen and Lubker, 1972). The tokens of each utterance type were aligned with reference to the onset of voicing for /u/, and were rectified, computer-sampled, integrated and averaged. Measurements of acoustic segment duration for the consonants were made from oscillographic displays of the speech waveform. Friction duration was measured for /s/ and closure duration, burst, and--where it was present--aspiration were measured for /t/.

RESULTS

The results for one of the two subjects are shown in Figure 1. There is a peak for the initial /p/ of each utterance, occurring roughly 500 msec before the onset of the vowel /u/. Rounding begins about 250 msec before the vowel /u/ for all eight utterance types. A similar picture is seen for the second subject, except that rounding begins about 175 msec before /u/.

These results are presented in more detail, with the acoustic measurements of consonant duration, in Table 1. As is well known, the durations of /s/ friction and /t/ closure and aspiration are shorter in clusters than in single-consonant, syllable-initial context. However, this compression is not sufficient to make cluster duration and single consonant duration identical--the clusters are somewhat longer for both subjects. When we turn to the EMG values we find a somewhat larger range of values, but no systematic relationship with consonant duration. Indeed, for one subject, KSH, EMG activity begins before consonant onset for three utterance types and after consonant onset for five utterance types.

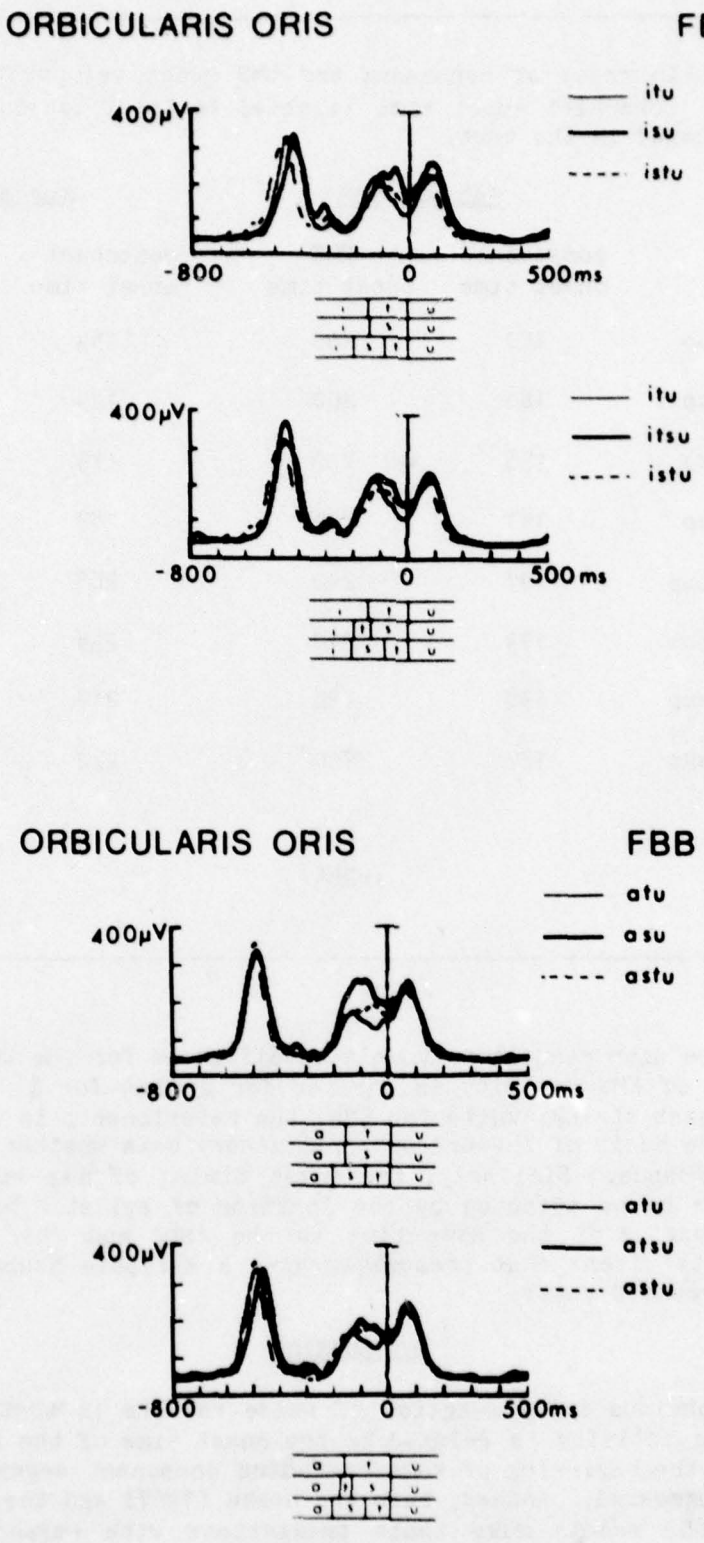


Figure 1: The results of the computer sampling and averaging of the data for one subject, FBB, for the four utterance types having /i/ as the first vowel in (a), and for the four utterance types having /a/ as the first vowel in (b). For convenience of comparison, two of the four averaged EMG traces are repeated in the two parts of (a) and (b).

TABLE 1: Time (in msec) of consonant and EMG onset before the rounded vowel /u/. Consonant onset time is equal to total consonant duration, as indicated in the text.

	Subject FBB		Subject KSH	
	consonant onset time	EMG onset time	consonant onset time	EMG onset time
pitup	160	250	153	210
patup	160	260	159	220
pisup	155	230	175	150
pasup	161	290	182	140
pistup	197	240	206	210
pastup	193	270	234	170
pitsup	198	230	214	140
patsup	187	250	222	180
		x=254		x=175

The picture with respect to vowels is different for the two subjects--for FBB, the onset of EMG activity is earlier for a than for i, when we compare the same consonant string, while for KSH, the relationship is variable. It is not clear on the basis of the present preliminary data whether this difference has any significance. Similarly, the onset timing of lip rounding activity does not appear to be affected by the location of syllable boundaries, since this activity begins at the same time in the /st/ and /ts/ items, and not later in the /ts/ items that presumably have a syllable boundary one segment closer to the rounded vowel.

DISCUSSION

The most obvious interpretation of these results is that the onset time of lip rounding activity is related to the onset time of the following vowel, rather than to the beginning of some preceding consonant segment, as previous studies have suggested. Indeed, both the Henke (1967) and the Kozhevnikov and Chistovich (1965) models make their predictions with respect to segmental units rather than to time.

In order to distinguish between segment-based and time-based hypotheses, it will be necessary to study longer consonant strings. In addition, we must collect EMG data from subjects other than the authors, whose data may reflect theoretical biases.

REFERENCES

Allen, G. D. and J. F. Lubker. (1972) New paint-on electrodes for surface electromyography. Journal of the Acoustical Society of America 52, 124(A).

Bell-Berti, F. and K. S. Harris. (1974) More on the motor organization of speech gestures. Haskins Laboratories Status Report on Speech Research SR-37/38, 73-78.

Benguerel, A.-P. and H. A. Cowan. (1974) Coarticulation of upper lip protrusion in French. Phonetica 30, 41-55.

Daniloff, R. G. and K. L. Moll. (1968) Coarticulation of lip-rounding. Journal of Speech and Hearing Research 11, 707-721.

Gay, T. J. (1975) Some electromyographic measures of coarticulation in VCV utterances. Haskins Laboratories Status Report on Speech Research SR-44, 137-145.

Henke, W. (1967) Preliminaries to speech synthesis based on an articulatory model. Conference Preprints; 1967 Conference on Speech Communication and Processing (Bedford, Mass.: Air Force Cambridge Research Laboratories), 170-177.

Kozhevnikov, V. A. and L. A. Chistovich. (1965) Rech', Artikulyatsiya, i Vosprivatiye. Trans. as Speech: Articulation and Perception, (1966). (Washington, D. C.: Joint Publications Research Service, 30), 543.

McLean, M. (1973) Forward coarticulation of velar movement at marked juncturaI boundaries. Journal of Speech and Hearing Research 16, 286-296.

Rapid vs. Rabid: A Catalogue of Acoustic Features That May Cue the Distinction*

Leigh Lisker+

ABSTRACT

In American English, initial /bdg/ often lack the acoustic feature taken as the defining feature of voiced stops; intervocalically before unstressed vowel /ptk/ lack aspiration, without which initial stops are not labeled "ptk." Initially the two categories differed in the timing of vocal fold adduction and onset of fold vibration; several acoustic cues, all tied to the VOT difference, have been studied. Medially there is also a difference in the management of the larynx, though it results in a phonetically simpler contrast, one of voicing with no accompanying difference in aspiration. Acoustically, however, the list of features that play, or might plausibly play a role is quite large. The word pair rapid-rabid, for example, might be affected by the following: 1) presence/absence of low frequency buzz during the closure interval; 2) duration of closure; 3) F_1 offset frequency before closure; 4) F_1 offset transition duration; 5) F_1 onset frequency following closure; 6) F_1 onset transition duration; 7) [æ] duration; 8) F_1 "cutback" before closure; 9) F_1 cutback following closure; 10) VOT cutback before closure; 11) VOT delay after closure; 12) F_0 contour before closure; 13) F_0 contour after closure; 14) amplitude of [ɨ] relative to [æ]; 15) decay time of glottal signal preceding closure; 16) intensity of burst following closure. Even if some of these should turn out to be perceptually negligible, enough of them surely have cue value to make it a formidable task to justify preferring an acoustic to an articulatory account of the distinction between the two English words.

If stop voicing continues to be a subject of lively interest to students of speech, it must be because it continues to provoke new questions or to refuse final answers to old ones. Perhaps this is because the stops of American English are not well chosen as the object of investigation whose purpose is to construct or test hypotheses concerning the perception of some single phonetic feature difference such as voicing. The phonetic differences between the American English phonemes /b/ and /p/, /d/ and /t/, and /g/ and /k/ are several, and we do violence to the internationally accepted definition of the term "voiced stop" if we call the phoneme set /bdg/ voiced, particular-

---

*This paper was presented at the 94th Meeting of the Acoustical Society of America, Miami Beach, December 1977.

+Also University of Pennsylvania.

Acknowledgment: The support of the National Institute of Child Health and Human Development is gratefully acknowledged.

ly if attention is focused on these categories in initial position. To refer to the acoustic features by which /bdg/ are distinguished from /ptk/ as cues to stop voicing is to produce a terminological muddle on two counts: 1) initial /bdg/ need not be, and often are not, produced with glottal pulsing before release, so that they may be voiceless in the sense of the word as used by the International Phonetic Association; and 2) the phonemes /bdg/ and /ptk/ are distinguished in different ways in different contexts, and constancy with respect to the phonetic feature of voicing is not a property of the phonologically contrasting sets.

Putting aside the matter of terminological purity and any question of cross-language validity, it does seem strange to construct hypotheses of stop voicing perception on the basis of the initial stops of American English, where the data derived from studying these events cannot represent the full range of phenomena that a theory of American English stop production and perception must encompass to be adequate. On the basis of traditional phonetic descriptions of these stops, and they seem to be taken seriously by most of us, it is hard to understand how any single detector, or detector-pair, could yield "outputs" that match the labeling behavior of English-speaking listeners. A detector that fires in response, let us say, to a periodically-excited transition following closure can tell the host-listener that a /bdg/ has occurred, but what inhibits a like response to medial /ptk/? The context-dependent nature of the phonetic differences between the two stop category sets of American English is a very old story for linguists, and it was in recognition of this fact that they were long disposed to find that the basis of the phonemic distinction was not one of voicing at all, but of something else they called "force of articulation." Voicing, it was sometimes said, is irrelevant to the contrast. More recently, a dimension of relative voice onset time (VOT) has been promoted as a measure by which to describe the difference between aspirated and unaspirated initial stops and the difference between voiced and voiceless medial stops, thereby avoiding recourse to the different and less accessible level of description at which something like "articulatory force" might be discovered and measured.

Medially in words before unstressed vowels American English /bdg/ are most often voiced, in the strict IPA sense, particularly where the signal both preceding and following closure is voiced. Just as commonly in that context, members of the other category set show voiceless closures. Here then is the place where the acoustic features that serve to distinguish the two sets of phonemes can be said to cue stop voicing. Oddly enough, although in this position the phonetic difference between the sets is considered to be smaller than it is in initial position, the number of acoustic pattern features whose manipulation may affect the labeling of a stoplike interval is much larger. However, it is only odd if we suppose that the number of phonetic features that differentiate the contrasting sets should determine the number of acoustic features we can isolate and manipulate to linguistic effect. Otherwise it is not particularly surprising: the initial stops cannot be cued by features preceding closure, nor are they usually cued by any feature of the closure interval itself. Of sixteen acoustic features that can, or can plausibly be supposed to serve as cues in the identification of members of a word pair like rapid-rabid, seven are to be found in the signal preceding the medial closure. These are: 1) duration of the [æ] vowel; 2) F_1 closing transition duration; 3) F_1 offset frequency before closure; 4) F_1 "cutback"

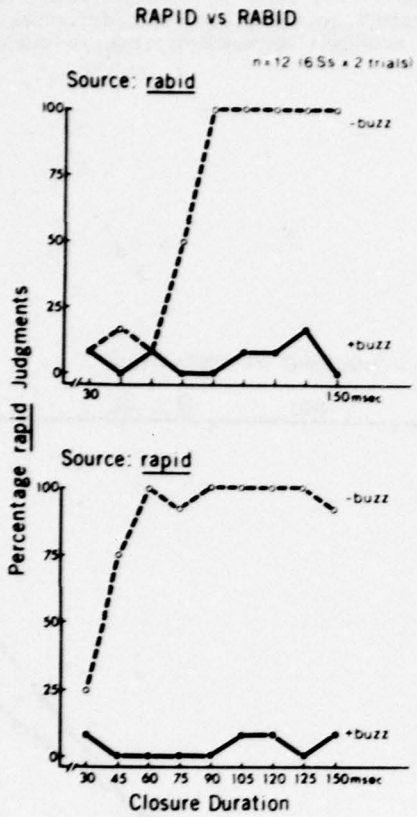


Figure 1: Responses of American listeners to stimuli derived from naturally produced tokens of the words rabid and rapid. Waveforms were edited to vary closure duration from 30 to 150 msec, in 15 msec steps. Closure intervals were either acoustically blank (-buzz) or entirely filled by laryngeally produced signal (+buzz) derived from the originally recorded token of rabid.

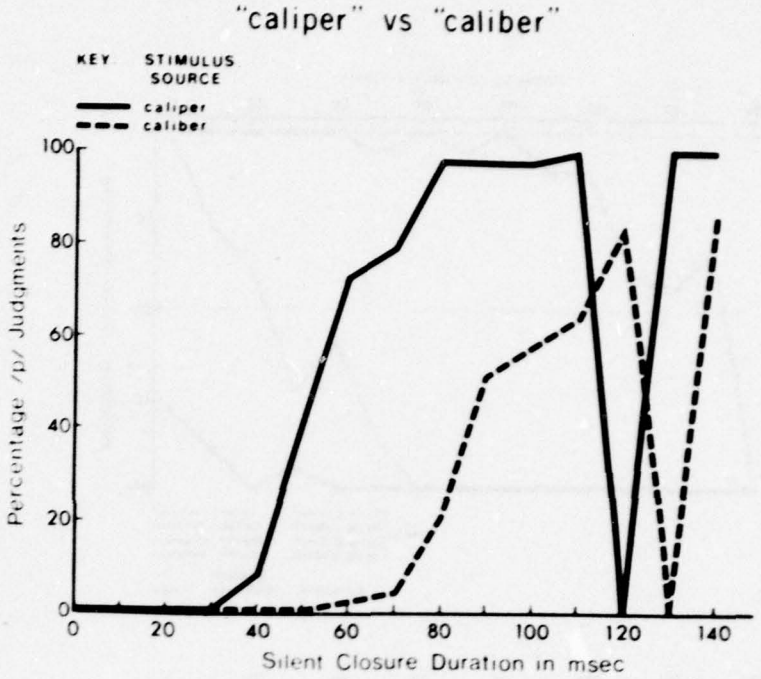


Figure 2: Responses of 12 listeners, all native speakers of American English, to stimuli derived by editing naturally produced tokens of caliber and caliper. All closure durations tested, except for 120 msec in the case of the caliper-derivatives and 130 msec for the caliber-derivatives, were acoustically blank.

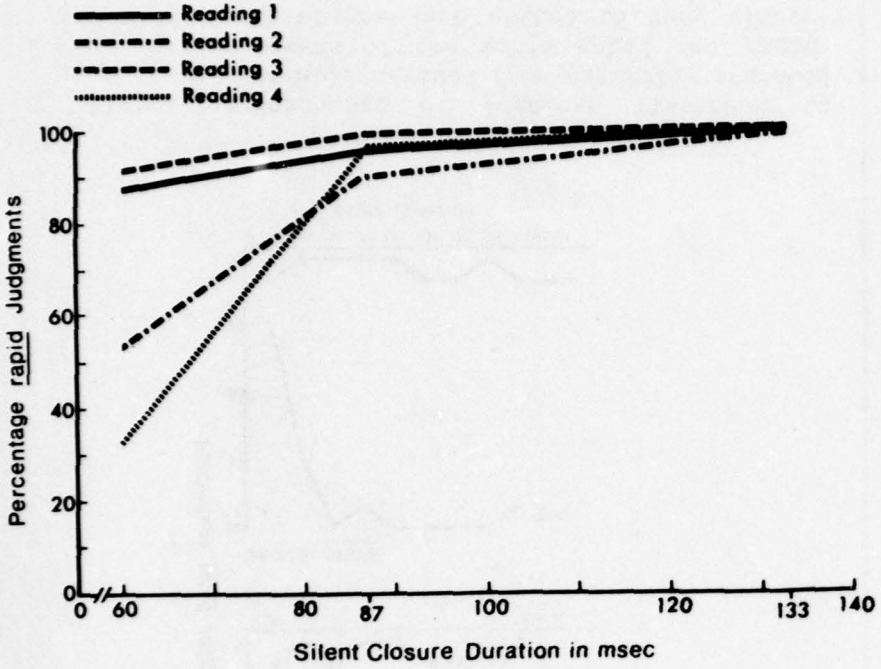


Figure 3: Responses of English-speaking listeners to stimuli derived from four naturally produced tokens of rapid. Closure durations were set at three values: 133 msec (an appropriate value for /p/), 87 msec (an appropriate /b/ value), and 60 msec (a value appropriate for /b/ and inappropriate for /p/).

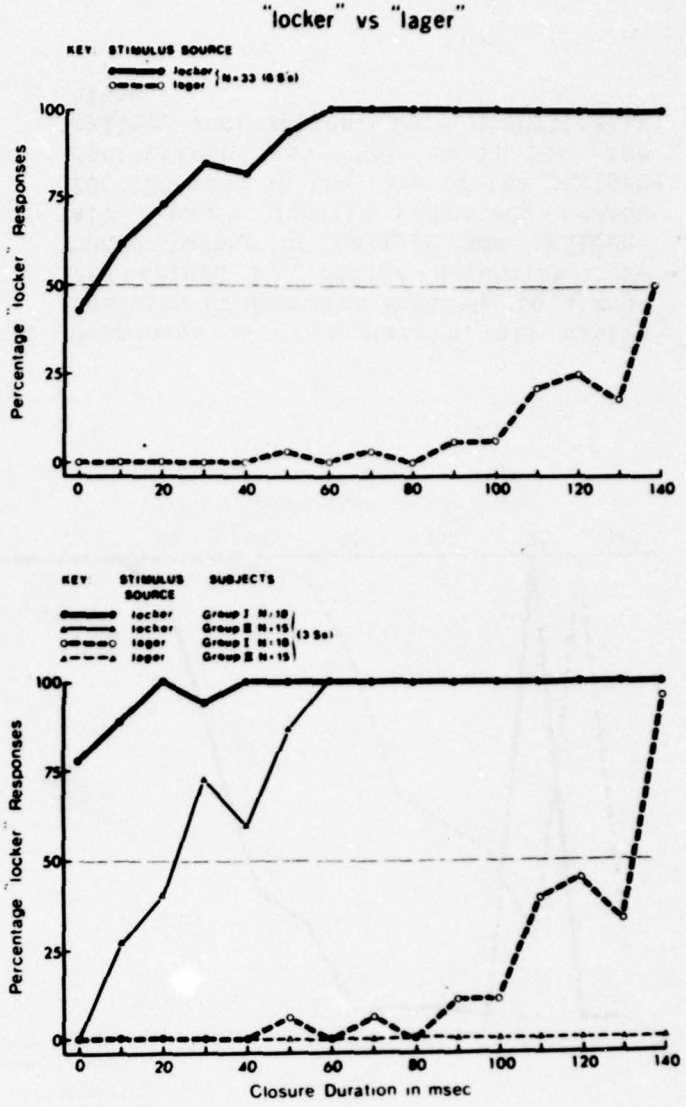


Figure 4: Responses of six English-speaking subjects to stimuli, all with acoustically silent closures, derived from naturally produced tokens of locker and lager. The data represented in upper and lower panels are the same; the lower display shows that the six subjects could be divided into two groups with rather different response patterns.

before closure; 5) timing of voice offset before closure; 6) F_0 contour before closure; 7) decay time of glottal signal at closure. Another feature, the intensity balance of the [æ] and [ɨ] vowels, which may affect stress-placement judgments and secondarily the evaluation of certain VOT values, is also restricted, obviously, to medial position. Two features of the closure interval itself, its acoustic nature and its duration, also play little or no role in initial position. The remaining six features, which are measures of the signal from release to following vowel, are simple mirror images of the features of the closing transition already mentioned.

Of the sixteen features that might affect the identification of a signal as rapid or rabid, it is probably true that none is indispensable and that several play no significant role in the perception of unedited naturally produced tokens of these words. If we are talking of acoustic cues, we do not limit ourselves to the perceptual evaluation of normally produced signals, and in particular we do not refrain from treating as independent variables features that are not independent in natural speech. If my reading of the experimental phonetic literature is correct, the conditions that an acoustic feature must satisfy in order to be called a "cue" do not involve a demonstrable conformity with nature; it is enough that patterns be devised so that manipulating the single feature effects a significant shift in listeners' word identification--say from "rapid" to "rabid." There is no requirement, it seems, that constant features of the test stimuli be copied from nature.

If the medial stops of rapid and rabid are correctly called voiceless and voiced respectively, it should suffice that they differ acoustically only over the interval corresponding to closure. In fact, we can delete the buzz from the closure of a naturally produced token of rabid to elicit "rapid" judgments. Moreover, a normal token of rapid with low-frequency buzz replacing its silent closure will be identified as "rapid," if the buzz is carefully tailored to its context. Figure 1 shows what listeners reported when tokens of the two words were provided with silent and buzzed closure intervals over various durations. Buzz-filled closures elicited mainly "rabid," with no effect of duration worth mentioning. Silent intervals were most often interpreted as /p/, although not when the closure was very brief. Thus it appears that buzz of any duration is incompatible with /p/, and that silent closures longer than about 100 msec preclude "rabid" responses.

Similar results were obtained by the same kind of editing of other naturally produced word pairs. In Figure 2, for example, data are the responses to stimuli derived from productions of the words caliber and caliper, where the closures were intended to be all silent, and the effect of varying these silent intervals was the object of interest. The unexpected results for caliper with 120 msec of silence and for caliber with 130 msec of silence is explained very simply; I made a mistake in fabricating these stimuli, providing them with buzzed instead of silent closures. However, this inadvertently obtained corroboration of the earlier finding, namely that the acoustic nature of the closure interval can determine listeners' responses, does not mean that we cannot also fail to obtain the same results with other natural tokens of words of this kind. When four different productions of rapid were tested with silent intervals copied from nature, they yielded the results shown in Figure 3. Two tokens were never heard as anything but "rapid," although the shortest duration tested was shorter than any /b/-

duration observed in the speaker's productions of rabid; one token of rapid was ambiguous for the shortest duration tested; only one was heard more often as "rabid" (66 percent) for this same duration. Possibly still shorter durations would have elicited more "rabid" judgments. Still other word pairs, locker-lager for example (Figure 4), resisted this treatment: locker with zero closure duration was still identified almost 50 percent as "locker," and lager with 140 msec of silent interval was reported equally often as "lager" and "locker." This result, possibly peculiar to the /g/-/k/ contrast, raises some questions for future discussion.

Measuring the cue values of the features preceding and following closure by presenting separately the first and second syllables of the natural speech disyllables gives data that are not easy to interpret; the part of rapid preceding closure is often not clearly heard to terminate in a /p/-closure, while the signal following closure is generally reported to begin with "b." It seems reasonable to suppose that in natural productions of a word like rapid, everything up to closure may be ambiguous so far as to whether a /b/ or a /p/ closure was executed, and that what follows closure is unambiguously /b/ if heard in isolation. Buzz in the intervening interval is decisive, while a short silent interval may or may not have the same outcome as buzz.

These results with natural speech can be described as less than clearcut, presumably because natural productions that are linguistic and even phonetic repetitions can differ in acoustic features that can be made to bear a greater perceptual load than they ordinarily have before mutilation of the signal. By pure synthesis it is possible to obtain results to indicate that silent intervals of any duration can be compatible with "b"; if the offset and onset frequencies of F_1 are very low, only "rabid" will be reported. In such patterns, the feature of [æ]-duration, which under other conditions may be decisive, can have no important effect on word identification.

In summary, in medial position the basis of the contrast between stop category sets is better identified with what happens in the closure interval than in other positions, but events elsewhere can make the closure interval alone insufficient to explain labeling behavior. The ensemble of features, spread over two syllables, shows a degree of disparity at the purely acoustic level that seems strange, given that they all affect the same phonetic judgment. However, they can all be referred back to a single crucial articulatory difference in the management of the larynx.

Acoustic Characteristics of Normal and Pathological Voices*

Steven B. Davis+

ABSTRACT

Recent studies have suggested that acoustic analysis of the pathological voice is a viable technique for early detection of laryngeal pathology or for clinical assessment of vocal improvement following voice therapy. Acoustic parameters may be extracted directly from speech or throat contact signals or indirectly from glottal or residue inverse filtered signals. These parameters characterize voice quality differences and may be used to discriminate between normal and pathological subjects. Such parameters may measure temporal variation, such as average pitch period or amplitude perturbation, or spectral variation, such as energy differences among frequency bands. This chapter discusses methods for the acoustic analysis of voices affected by laryngeal pathology, and procedures for determining acoustic parameters applicable to screening and clinical assessment. These methods use digital computer techniques for voice analysis to extract acoustic measures of vocal function from the speech signal. A brief review provides information on some auditory and visual methods for diagnosing laryngeal pathology. Then the vocal fold movement is related to the acoustic output on which the acoustic methods are based, and the theoretical bases and results of several methods are compared. Finally, acoustic parameters and representative waveforms based on inverse filtered speech are used in a Voice Profile to assess early cases of pathology and to monitor progress during voice therapy.

INTRODUCTION

In recent years, researchers in such fields as laryngology and speech science have become increasingly interested in the acoustic characteristics of normal and pathological voices (Murry, 1975; Davis, 1975, 1976b; Hiki, Imaizumi, Hirano, Matsushita and Kakita, 1976). One reason for this trend is

---

*To appear in Speech and Language: Research and Theory, ed. by N. K. Lass. (New York: Academic Press).

+Also Speech Communications Research Laboratory, Inc., Santa Barbara, California.

Acknowledgement: The preparation of this chapter was supported by National Institutes of Health Grants NS 13309-01A1 to Speech Communications Research Laboratory, Inc. and NS 13870 to Haskins Laboratories and the Voice Foundation, New York.

that acoustic methods have the potential to provide quantitative techniques for the clinical assessment of laryngeal function. The desire for an objective method for analysis of the pathological voice was expressed by the 1973 Conference on Early Detection of Laryngeal Pathology (Moore, 1973, p. 6):

> "The otologist and audiologist can employ standardized audiometric tests to evaluate hearing, the cardiologist has access to electrocardiograms to evaluate heart function, but the laryngologist and speech pathologist have no comparable aids that can be used in the clinical setting."

There are several methods currently used in laryngeal research and diagnosis, for example, laryngoscopy, stroboscopy, thermography, electromyography, pneumotachography, glottography and acoustic analysis, but acoustic analysis appears to have an advantage for routine clinical evaluation of laryngeal function, for example, during a program of voice therapy, because of its nonintrusive nature and its potential for providing quantitative data with reasonable expenditures of analysis time.

In addition to its potential value in the clinical assessment of laryngeal function, a sensitive automatic acoustic technique could be used to screen individuals for early cases of laryngeal pathology. Moore (1973) has indicated that none of the other techniques are useful for screening, and are only applied if an individual specifically seeks aid. The development of portable instrumentation for acoustic analysis would lead to programs similar to audiometric testing in schools, industry, etc. Such instrumentation could have large benefits in terms of overall health maintenance.

This chapter discusses methods for the acoustic analysis of voices affected by laryngeal pathology and procedures for determining acoustic parameters applicable to screening and clinical assessment. These methods use digital computer techniques for voice analysis (developed originally to increase the efficiency of speech transmission systems) to extract acoustic measures of vocal function from the speech signal. A brief review provides information on some auditory and visual methods for diagnosing laryngeal pathology. Then the vocal fold movement is related to the acoustic output on which the acoustic methods are based, and the theoretical bases and results of several methods are compared, indicating the difficulties requiring future research.

AUDITORY AND VISUAL METHODS

Historically, physicians have relied on two basic techniques in the diagnosis of pathological conditions of the larynx: 1) listening to the voice, and 2) viewing the larynx with a mirror or laryngoscope. Since laryngeal diseases are often accompanied by voice quality changes, simple listening tests sometimes give useful information. The principal criticisms of listening tests are their subjectivity (that is, even experienced laryngologists may offer different diagnoses for the same patient), and the related problem of the lack of quantitative standards.

Visual observations allow a physician to substantiate auditory evaluations. In indirect laryngoscopy, the larynx is viewed via a mirror inserted

into the back of the mouth. The gross structure and movements of the vocal folds are observed; however, the amount of detail made available by indirect laryngoscopy is limited because the field of view is small and the distance from the larynx is relatively long. Pathologies beneath the vocal folds frequently can be overlooked because only the superior surfaces of the larynx may be visualized from above. Also, the rapid vibratory motions of the vocal folds cannot be observed with indirect laryngoscopy. However, high quality photographic records of the laryngeal image exposed by advanced laryngoscopes using fiberoptics (Sawashima and Hirose, 1968; Gould, 1973) have enhanced the clinical value of laryngoscopy.

In direct laryngoscopy a viewing tube is inserted directly into the larynx. Direct laryngoscopy is not used on a wide scale because it is a medical procedure requiring anaesthesia, it disturbs the positioning and function of the articulatory structures, and it is uncomfortable for the patient. Consequently, its application usually is confined to diagnosis and verification during the later stages of a laryngeal disease and to surgical situations (Koike, 1976).

Stroboscopic laryngoscopy combines a laryngeal mirror and a high flash-rate stroboscope to give either a stationary or slowly moving image of the vocal folds (Moore, 1938; Schönhärl, 1960; van den Berg, 1962). However, the image is a composite of many cycles of the vibration, and fine details of the individual periods are not observed.

The detailed movement of the vocal folds during individual periods can be seen, however, if a high-speed motion picture camera, a special light source and a laryngeal mirror are used (Farnsworth, 1940; Koike and Takahashi, 1971). Compared to stroboscopic laryngoscopy, this high-speed technique captures all vibratory behavior in full detail. Studies based on high-speed motion pictures have demonstrated various vibratory patterns in patients with laryngeal pathology (Timcke, von Leden and Moore, 1958, 1959; von Leden, Moore and Timcke, 1960). In one study (Moore, 1968), high-speed motion pictures of pathological vocal folds were synchronized with acoustic recordings of the voice. Moore demonstrated complex and irregular vibratory patterns resulting in complicated changes in glottal width. He also noted the independence of each fold as a vibrator and suggested that this independence should be considered as a cause of hoarseness. This independence was confirmed by Koike and Hirano (1973).

High-speed motion picture analysis is an important tool in basic voice and speech research, but there are several limitations for large-scale clinical applications. The most significant limitation is the time-consuming task of frame-by-frame data analysis. Even with the aid of digital computer programs designed to simplify the measurement process (Ramsey, 1964; Soron, 1967; Koike and Takahashi, 1971; Hayden and Koike, 1972; Tanabe, Kitajima, Gould and Lambiase, 1975; Hildebrand[1]), it is not a trivial task to record the

---

[1]Hildebrand, B. (1976) Vibratory patterns of the human vocal folds during variations in frequency and intensity, Doctoral Dissertation, University of Florida, Gainesville.

glottal widths or areas for a large number of frames. One preliminary study (Davis, 1976a) applied digital image processing to automatically extract the glottal area information from successive picture frames, but more research is needed before the method can become clinically useful. Nevertheless, the information obtained from high-speed motion pictures that are synchronized with the voice provides a basis for gaining a better understanding of the relationship between the acoustic signal and the physiological function of the larynx. Some of the parameters that can be measured from successive high-speed motion picture frames are the open quotient (OQ), which is the ratio of the open time of the glottis to the total time of one vibratory cycle, and the speed quotient (SQ), which is the ratio of the time of vocal fold abductory movement to the time of adductory movement. These parameters are important for assessing the vibratory behavior of the vocal folds (Timcke, von Leden and Moore, 1958, 1959; von Leden, Moore and Timcke, 1960; Hildebrand, see footnote 1).

ACOUSTIC SYMPTOMS OF LARYNGEAL PATHOLOGY

Acoustic analysis of the voice is more objective than auditory methods for screening or voice therapy assessment (Koike, 1976). The validity of the acoustic approach, however, rests on the complex relationship between the physiological source function and the concomitant speech signal. A laryngeal pathology, such as tumors or paralysis, generally produces assymetrical changes in the mass, elasticity and tension of the vocal folds, leading to deviant vibration. Also, weakness or paralysis of respiratory muscles may cause insufficient subglottal pressure, thus changing the aerodynamic forces acting on the vocal folds and hence their vibratory pattern. The subglottal airstream is modulated by this unbalanced vocal fold movement. Irregular air pulses emerge from the larynx, propagate through the pharynx and oral and nasal cavities, and radiate from the mouth and nose. The resultant acoustic signal is thus affected by a physiological disturbance in the larynx, and the acoustic signal may be used to measure the disturbance.

The primary acoustic symptoms of laryngeal pathology are a change in fundamental frequency, voice intensity or voice quality. These symptoms are indicative of a multitude of organic diseases and functional disturbances (Zemlin, 1968; Moore, 1971), and the nature of these symptoms will vary for each patient and at each stage of pathological involvement.

Fundamental Frequency

The fundamental frequency of a voiced sound is a function of the mass, elasticity, compliance and length of the vocal folds. It also depends somewhat on the subglottal pressure and the configuration (acoustic load) of the vocal tract. An assessment of whether a patient has adequate frequency regulation usually involves a judgment by a trained listener as to whether the fundamental frequency is too high or too low when compared to voices of persons of similar age, sex and body size. If the fundamental frequency is judged to be too high, the voice may sound "shrill" or "screechy." In functional disorders involving high fundamental frequency, the vocal folds tend to become abused at the point of maximum displacement. Vocal abuse may produce laryngitis, lead to the development of nodules, or worsen an existing pathology. Organic causes for high fundamental frequency include a laryngeal

web, asymmetrical structures or failure of a male larynx to develop to a normal size. A fundamental frequency which is judged to be too low may sound "harsh," "hoarse," "husky" or "rough." Low fundamental frequency stemming from functional vocal abuse may lead to contact ulcers, although the precise etiology of contact ulcers is unclear. Organic causes may be virilization, tumors or other enlargements that increase the mass of the folds, or nerve paralysis that decreases the elasticity and compliance of one or both folds (Luchsinger and Arnold, 1965).

Vocal Intensity

Vocal intensity is a monotonically increasing function of the SQ and the air flow through the glottis, and it is a monotonically decreasing function of the OQ (Timcke, von Leden and Moore, 1958). Excessive vocal intensity usually is functional in origin. If it is also accompanied by high fundamental frequency, the voice may sound "shrill" or "screechy," and if it is also accompanied by low fundamental frequency, the voice may sound "hoarse." When excessive vocal intensity is coupled with excessively high or low fundamental frequency, the severity of a pathology may increase. Weak voices generally have organic causes, and the etiology may be attributed to insufficient subglottal pressure caused by paralysis of the respiratory muscles, or to poor vocal fold movement caused by muscle weakness or laryngeal paralysis (Luchsinger and Arnold, 1965).

Vocal Quality

A degradation in voice quality, generally categorized as "hoarseness," is often the first and sometimes the sole symptom of laryngeal disease. This change, because of its initial presence, drew the early attention of several researchers (Jackson and Jackson, 1937; Frank, 1940; Arnold, 1955; Palmer, 1959; Bowler, 1964). However, these studies were perceptual, and there was a multitude of concepts and terms. One review (Perkins, 1971) compared nine studies that assess quality defects and listed twenty-seven terms used to describe those defects. Only "hoarseness" and "nasality" appear in all studies, and only ten other terms appear in more than one study. Of these ten terms, six are common to four studies: "breathy," "harsh," "strident," "denasal," "husky," and "metallic." Other terms are "screechy," "guttural," "throaty," "strained," "shrill" and "intense." This review demonstrates that there is little agreement among researchers on describing voice quality, and there is a proliferation of descriptive terms.

This complex terminology led some researchers to attempt to identify the factors that are involved in listener judgments of pathological voice quality (Isshiki, 1966; Isshiki, Okamura, Tanabe and Morimoto, 1969; Takahashi and Koike, 1975; Fritzell, Hammarberg and Wedin, 1977). Using a technique based on the semantic differential (Osgood, Suci and Tannenbaum, 1957), Isshiki (1966) suggested that the factors that operate in listener judgments are multidimensional, and he identified four major independent axes corresponding approximately to "roughness," "breathiness," "lack of power," and "normalcy." Using principal components analysis, Fritzell, Hammarberg and Wedin (1977) identified five factors as "stable-unstable," "breathy-overtight," "hypo-hyperkinetic," "light-course" and "head-chest register."

In summary, vocal quality is a difficult parameter to assess, and unlike fundamental frequency or vocal intensity, no reliable physiological descriptions or measurements have been established. In general, vocal quality is determined on the basis of vocal fold vibration and vocal tract resonance (Luchsinger and Arnold, 1965). Furthermore, asymmetrical vibrations are a typical indication of a vocal quality defect. Such variations affect the fundamental frequency, SQ and OQ, and may be attributed to changes in the mass, elasticity, compliance or length of one or both folds (von Leden, Moore and Timcke, 1960). Vocal phonatory defects range from "aphonia" and "breathiness" to "hoarseness" and "rough hoarseness," and the organic causes include paralysis, weak muscles, extraneous masses, excessive mucous and loss of tissue (Luchsinger and Arnold, 1965).

ACOUSTIC TECHNIQUES FOR VOICE ANALYSIS

Acoustic techniques for voice analysis are based on the source of the signal and the method of analysis. One source is direct signals, for example, radiated sound pressure and throat contact signals. The other source is indirect signals, for example, glottal or residue signals derived using inverse filtering techniques. For each of these sources, the signals may be analyzed in the time domain, for example, to determine mean fundamental frequency or mean perturbation, or in the frequency domain, for example, to determine long-term average spectral slope or energy distribution. The following sections will discuss the advantages and limitations of each signal source and analysis method.

Direct Signals

The radiated sound pressure waveform is the most readily available signal for acoustic analysis, but its usefulness for assessing laryngeal function is limited. The production of a steady vowel sound is controlled by the glottal source, which will be affected by laryngeal pathology, and the supraglottal structure, whose resonance characteristics will presumably be unaffected by laryngeal pathology. Therefore, acoustic measures of a laryngeal disorder from a sound pressure waveform are affected by a normal supraglottal structure. The effects of the supraglottal structure do not significantly hinder the detection or assessment of severe laryngeal disorders. However, at an early stage of pathology, or at a late stage of recovery, the supraglottal structure probably masks some of the important acoustic attributes of the pathology. A throat contact microphone is sometimes used to avoid supraglottal effects (Koike, 1969), but information from throat contact signals is limited by the low-pass filtering action of intervening tissues, and also because the throat signal is sensitive to microphone placement.

Indirect Signals

A voiced sound such as a sustained vowel may be simply modeled as the sound pressure waveform resulting from the excitation by a periodic source (corresponding to the glottis) of an acoustic tube (corresponding to the vocal tract and lips). The technique of inverse filtering applied to the sound pressure waveform can remove the effects of the acoustic tube, and the resulting signal approximates the periodic source. If the supraglottal structure is relatively unaffected by laryngeal pathology, and the source of a

voice change is the glottis, then this periodic source approximation contains sufficient information to analyze the acoustic effects of the pathology. Measures based on an inverse-filtered speech signal are not affected by the supraglottal structure, and are potentially more informative than measures based on an unfiltered speech signal.

There are two inverse filtering methods that are used to obtain acoustic information about the glottal source. The first method is glottal inverse filtering. In this procedure, the inverse of the lip radiation and vocal tract spectral contributions is used to estimate the glottal volume velocity waveform as a function of time (Miller, 1959; Holmes, 1962; Lindqvist, 1965; Takasugi and Suzuki, 1970; Rothenberg, 1973). The theoretical starting point for glottal inverse filtering is the linear voiced speech production model (Fant, 1960; Flanagan, 1972). Early glottal inverse filtering studies used analog techniques, and usually only eliminated the first and second formants. Later studies employed digital computer techniques to process the speech signal in an interactive manner. The formant frequencies and bandwidths were estimated and used to adjust a glottal inverse filter until the expected waveform (roughly triangular in shape followed by a zero portion) appeared (Hiki et al., 1976). However, the overall appearance of the resulting waveform generally did not satisfy intuitive concepts of vocal fold closure. The linear predictive technique of pitch-synchronous glottal inverse filtering produces acceptable waveforms without estimation of the formants, but the point of glottal closure for each pitch period must be located by visual inspection (Wong and Markel, 1976).

Another method of glottal inverse filtering eliminates the vocal tract resonance by having a subject speak into a fairly long reflectionless tube (approximately 2.5 cm x 100 cm) (Sondhi, 1975). The resulting waveform is closely correlated with the glottal volume velocity waveform. The method is subjective, however, since each speaker requires a different tube (matched to their vocal tract) for the best results.

Although some of these techniques have produced good estimates of the glottal signal, none of these techniques is completely automated. Glottal inverse filtering can become a useful clinical tool only when the technique requires no user-interaction.

The second indirect method for obtaining information about laryngeal activity is residue inverse filtering. This technique is based on a linear model of speech production, (Atal and Hanauer, 1971; Markel and Gray, 1976; Davis, 1976b). The residue inverse filter is the inverse of the estimated lip radiation, vocal tract and glottal shaping spectral contributions to the speech signal. The result of filtering the speech signal with the residue inverse filter is the residue signal. This signal is an estimate of a periodic source function for an all-pole speech production model, and it exhibits strong peaks at the start of each pitch period and quasirandom noise between the pitch period peaks. Koike and Markel (1975) used the residue signal in a qualitative analysis of normal and pathological voices. They indicated that for some intermediate and advanced cases of laryngeal pathology, the residue signal did not appear to convey more information about the vocal disorder than was already apparent in the speech signal. However, for some early cases, Koike and Markel claimed that the residue signal showed

qualitative evidence of pathology even though the unfiltered speech signal showed no such evidence. Subsequently, Davis (1976b) substantiated these claims by developing quantitative measures based on the residue signal, and verified the hypothesis that more acoustic information about a pathology is conveyed by the (indirect) residue signal than the (direct) speech signal.

Comparison of Direct and Indirect Signals

Figure 1 depicts the sound pressure waveform (A), the glottal inverse-filtered signal (B), and the residue inverse-filtered signal (C) for a vowel segment. The inverse filter that produces the residue signal differs from the inverse filter that produces the glottal signal by the addition of several low frequency poles. Thus the glottal signal is equivalent to a low-pass filtered version of the residue signal. In practice, the residue signal is considerably easier to obtain than the glottal signal, since glottal inverse filtering has not been automated. As noted above, the glottal inverse filter coefficients must either be determined from the estimated formant frequencies and bandwidths (Hiki et al., 1976), or the point of glottal closure must be marked for automatic evaluation of the formants (Wong and Markel, 1976). In addition, any low frequency phase distortion will prevent accurate approximation of the glottal signal (especially its closed interval), so the speech signal must be recorded on equipment having a good low frequency phase response even under 100 Hz (Holmes, 1975). A standard audio tape recorder has a poor low frequency phase characteristic, and it is extremely difficult to estimate the glottal signal from a speech signal recorded on such equipment. An FM tape recorder avoids phase problems, but such tape recorders are generally not available in a clinic. In contrast, the residue inverse filter uses a phase-insensitive autocorrelation method to match the overall spectral envelope of the vocal tract, glottal shaping and lip radiation. There is no need for visual inspection of formants, marking of pitch periods, or controlled recording conditions.

In a theoretical sense, the glottal signal has an important advantage over the residue signal, since the glottal signal is a good approximation to the glottal volume velocity waveform (Figure 1), while the residue signal is not directly related to any physically-observable signal. In a practical sense, however, the residue signal may be more easily obtained, and hence has greater potential value than the glottal signal for the clinical assessment of pathological voices.

Time-Domain Acoustic Parameters

High-speed motion pictures of pathological vocal folds have revealed that there frequently are irregular vibratory patterns (von Leden, Moore and Timcke, 1960). Pitch period perturbation measures (Lieberman, 1961, 1963; Smith and Lieberman, 1964; Koike, 1967, 1973; Hiki, Sugawara and Oizumi, 1968; Crystal and Jackson, 1970; Hecker and Kreul, 1971; Kitajima, Tanabe and Isshiki, 1975; Takahashi and Koike, 1975; Davis, 1976b) and amplitude perturbation measures (Koike, 1969; Takahashi and Koike, 1975; Davis, 1976b) from acoustic signals are different between pathological and normal speakers. Two psychoacoustic studies (Wendahl, 1963, 1966) involved the synthesis of speech with pitch period and amplitude perturbations. The degree of perturbation in

the sounds was closely correlated with subjective judgments of degree of "roughness."

Measures based on pitch period and amplitude perturbations have been formulated in several different ways. The first of these measures is the "pitch perturbation factor" (Lieberman, 1961, 1963). This parameter is defined as the relative frequency of pitch period perturbations larger than 0.5 msecs occurring in a steady vowel sound. A pitch period perturbation is defined as the time difference between the durations of successive pitch periods in the speech signal. Lieberman showed that pathological voices generally have larger perturbation factors than normal voices with comparable fundamental frequencies, and that the perturbation factor is sensitive to the size and location of growths in the larynx.

A second perturbation measure is the "relative average perturbation" (Koike, 1973). This parameter differs from Lieberman's pitch perturbation factor in several respects. Koike observed that steady vowel sounds may normally exhibit slow and relatively smooth changes in pitch period, and he measured rapid perturbations from a smoothed trend line. In addition, Koike suggested normalizing the pitch period perturbation measure by dividing it by the average pitch period. Lastly, Koike suggested that the throat contact signal is a better indicator of laryngeal aperiodicity than the speech signal itself because the effects of the supraglottal structure on the speech sound are reduced. In consideration of these observations, Koike defined the relative average perturbation (RAP) as:

$$RAP = \frac{\frac{1}{N-2}\sum_{i=2}^{N-1}\left|\frac{P(i-1)+P(i)+P(i+1)}{3} - P(i)\right|}{\frac{1}{N}\sum_{i=1}^{N}P(i)}, \qquad (1)$$

where $P(i)$, $i = 1,2,\ldots,N$, denote the successive pitch periods. The numerator is the average perturbation measured for each pitch period smoothed by a three-point averaging window, and the denominator is the average pitch period. Koike concluded that the RAP of normal and pathological voices have different ranges, and that the RAP varies significantly between patients with neoplasms and patients with unilateral paralysis.

The basis for a third measure of perturbation is the quasiperiodic amplitude modulation observed in the steady vowel sounds of pathological speakers (Koike, 1969). Koike computed the serial correlation coefficients (correlogram) for the time series of amplitude values at each pitch period peak. He found that the correlograms of normal and pathological speakers are generally distinguishable from one another. The correlograms for speakers with laryngeal tumors usually show significant correlation peaks at lags between three and twelve fundamental periods, and the correlograms for speakers with laryngeal paralysis show no such peaks. Koike concluded that it might be possible to develop methods for differential evaluation of laryngeal pathologies based on information in the amplitude envelope of steady vowel sounds.

In a fourth approach, Kitajima, Tanabe and Isshiki (1975) defined a fundamental frequency (F_0) perturbation measure as the average of 100 successive F_0 perturbations from an all-voiced phrase. An F_0 perturbation is the difference between a measured F_0 and a five-point weighted (in a least-squares sense) average centered around the measured value. A semitone scale (relative to 16.35 Hz or four octaves below middle C) was used, since auditory perception of F_0 is approximately proportional to the logarithm of frequency. They found that normal female voices exhibit larger F_0 perturbation measures than normal male voices, which would be expected since the female F_0 is generally higher than the male. They also found that male voices affected by laryngeal cancer are distinguishable from normal male and female voices by using the F_0 perturbation measure.

Further advances in perturbation measures were made when Takahashi and Koike (1975) combined Koike's earlier results into two time-domain measures of the pathological voice based upon signals obtained from the throat contact microphone; the frequency perturbation quotient (FPQ) and the amplitude perturbation quotient (APQ). The FPQ is defined analogous to the RAP, but the instantaneous F_0 (defined as the reciprocal of the pitch period) is substituted for the pitch period. The APQ is also defined analogous to the RAP, but now peak amplitude values for each pitch period and an eleven-point, rather than a three-point, averaging window are used. Takahashi and Koike found that the APQ made a significant contribution in a principal components analysis of voice quality factors, and that although the FPQ was significantly correlated with the APQ, the FPQ did not make a significant contribution in the principal components analysis.

Davis (1976b) defined a pitch period perturbation quotient (PPQ) and an amplitude perturbation quotient (APQ), which respectively were analogous to Koike's RAP and Takahashi and Koike's APQ. However, there were several differences in the acoustic definitions. Rather than using fixed three-point or eleven-point averaging windows, Davis systematically investigated the benefit of changing the window size, and found that five-point averaging windows for the PPQ and APQ produce the best perturbation measures for normal-pathological discrimination. Davis also found that perturbation measures based on the residue signal are better for discrimination than those based on the speech signal, but he did not attempt any comparisons with the throat contact or glottal signals used by other researchers.

Davis also developed a completely automated procedure for extracting pitch period and amplitude sequences. Once the residue signal is obtained, the extraction procedure uses a peak picking algorithm that finds the significant positive and negative excursions of the signal (Figure 2). The APQs for the positive and negative amplitude sequences are calculated, and the smallest APQ is choosen. The PPQ is then found from the pitch period sequence corresponding to the smallest APQ.

Davis developed two additional time-domain acoustic measures of laryngeal pathology. One measure was based on the observation that the signal-to-noise ratio of the residue signal is a good cue for normal-pathological discrimination (Koike and Markel, 1975; Davis, 1976b). The "signal" in this case is the sequence of spikes spaced at pitch period intervals, and the "noise" is the quasi-random energy between the spikes. Koike and Markel suggested that one

measure of signal-to-noise ratio might be the average of the peak energy for each period divided by the noise energy in the last half of the period for successive pitch periods in a specified interval, but they did not actually attempt any quantitative measurements.

For a residue signal from a normal speaker, the separation of signal from noise for each pitch period is straightforward, and a measure such as the one described by Koike and Markel would suffice. However, for a residue signal from a pathological speaker, the pitch peak is not always distinct from the noise, and it would be very difficult to implement such a measure. The appearance of more noise in pathological cases, and less noise in normal cases, suggested to Davis that the amplitude distribution of the residue signal would be useful for a statistical measure of the signal-to-noise ratio. Figure 3 shows normal and pathological residue signals and the corresponding amplitude distributions. The distribution for the normal speaker is taller and narrower than the distribution for the pathological speaker.

The shape of these distributions may be quantified by a statistical measure called the coefficient of excess (EX) (Cramer, 1958). This coefficient is defined as the ratio of the fourth moment of a distribution to the square of the second moment of the distribution. The EX is zero for a Gaussian distribution. That is,

$$EX = \frac{E\{(x - \bar{x})^4\}}{E\{(x - \bar{x})^2\}^2} - 3 \quad , \tag{2}$$

where

$$E\{(x - \bar{x})^k\} = \frac{1}{N} \sum_{i=0}^{N-1} [x(i) - \bar{x}]^k \quad , \tag{3}$$

and

$$\bar{x} = \frac{1}{N} \sum_{i=0}^{N-1} x(i) \quad , \tag{4}$$

for a signal x(i), i = 0,1,...,N-1. The EX is positive if the distribution is taller and narrower than a Gaussian distribution and negative if the distribution is shorter and wider. Measurements from numerous speakers substantiate the correlation between the values of the EX and a judgment of the residue signal-to-noise ratio (Davis, 1976b).

The other time-domain acoustic parameter developed by Davis is based on the amount of voicing, or the strength of F_0 during a sustained vowel sound. The detection of F_0 is important in almost any analysis or synthesis study involving speech. In synthesis experiments, for example, the voiced-unvoiced decision is based on the presence or absence of F_0 and is used to determine

whether impulses or noise should be used for the source excitation. One of the oldest digital methods for detecting F_0 is autocorrelation analysis (Markel, 1973). Figure 4 shows that a periodic signal, for example, a vowel, exhibits a peak in the autocorrelation function of the residue signal at a duration corresponding to the period. The reciprocal of the period yields the fundamental frequency. Alternately, an aperiodic signal, for example, an unvoiced fricative, shows no pitch period peak. From Figure 4, it is evident that the residue signal is a better indicator of the autocorrelation peak than the speech signal. Davis (1976b) defined a time-domain acoustic parameter called the pitch amplitude (PA) as the value of the pitch period peak in the residue signal autocorrelation function. The PA is high for vowels, small for voiced fricatives, and zero for unvoiced fricatives.

If a given sound is known to be voiced, then the PA becomes a measure of voicing. Voiced sounds from normal speakers have a clearly defined pitch period and the PA is high. In these cases, there is strong periodicity in the glottal volume velocity and area waveforms associated with symmetrical vocal fold movements. Alternatively, "breathy" voiced sounds from pathological speakers are acoustically analogous at the source level to unvoiced sounds from normal speakers. The PA is low or not measurable, the speech sounds have weak periodicity, and there is a significant increase in the amount of noise which is heard.

Frequency-Domain Acoustic Parameters

As an alternative to time-domain analysis, frequency-domain analysis provides a different set of acoustic features. A common instrument for frequency analysis of speech is the sound spectrograph, which analyzes the spectral energy distribution of a short speech segment (generally less than three seconds) by filtering the sound with a tracking bandpass filter. The output is a time versus frequency graph of the sound, with spectral energy indicated by intensity. The formants and F_0 of a steady vowel are readily visualized in a spectrogram.

Several spectrographic studies show that there are differences between the spectra of pathological voices and the spectra of normal voices (Winckel, 1952, 1954; Nessel, 1960; Yanagihara, 1967a, 1967b; Gould[2]). The higher frequency harmonics of steady pathological vowels are attenuated in comparison with their normal counterparts. The loss of high frequency harmonics may be caused by changes in the OQ or SQ. In particular, if there is no glottal closure (that is, the OQ is equal to unity), the higher harmonics are sharply attenuated. Spectral noise components may originate in turbulent air flow resulting from incomplete glottal closure or irregular vocal fold vibration (Flanagan, 1958). These spectral noise components are distributed over the spectrum in varying degrees, and the extent of the distribution depends on the severity of the disease. The presence of spectral noise contributes to "hoarseness," which is the first symptom of numerous pathologies; some laryngologists use spectrograms in assessing the degree of and recovery from vocal fold disorders (Rontal, 1975; Gould, 1975). The results of these

---

[2]Gould, W. J. (1976): personal communication.

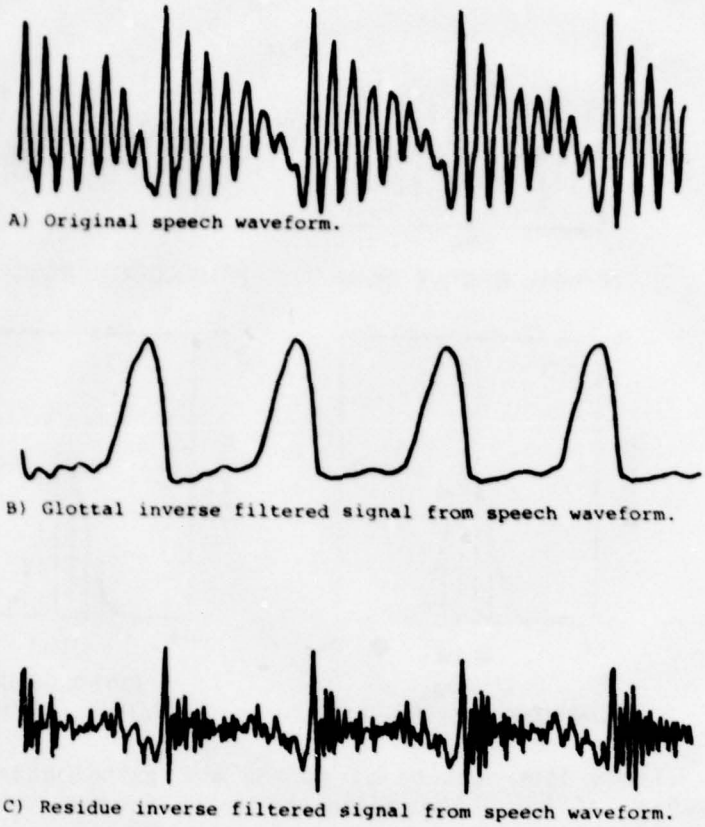


Figure 1: Comparison of speech, glottal and residue signals for /a/. The glottal signal is closely correlated with the physiological glottal volume velocity waveform (Davis, 1976b).

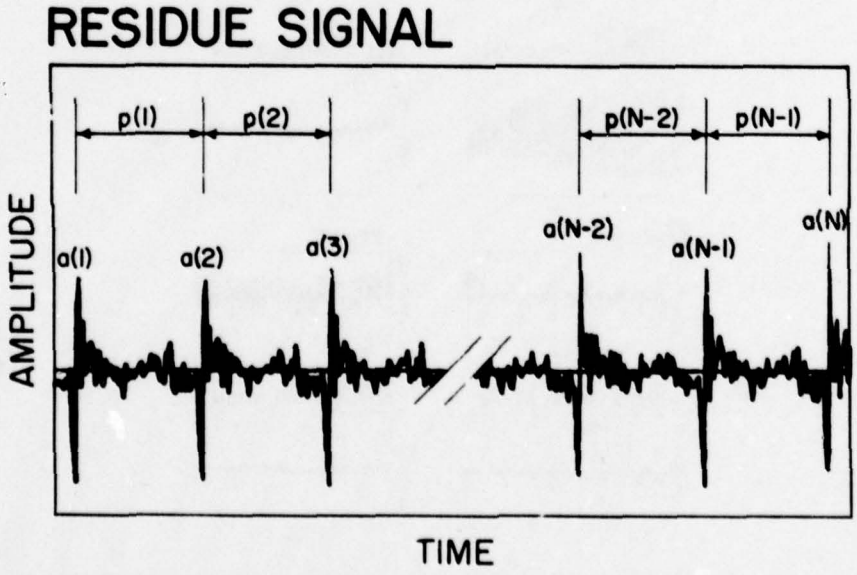


Figure 2: Extraction of pitch period and amplitude sequences from the residue signal (Davis, 1976b).

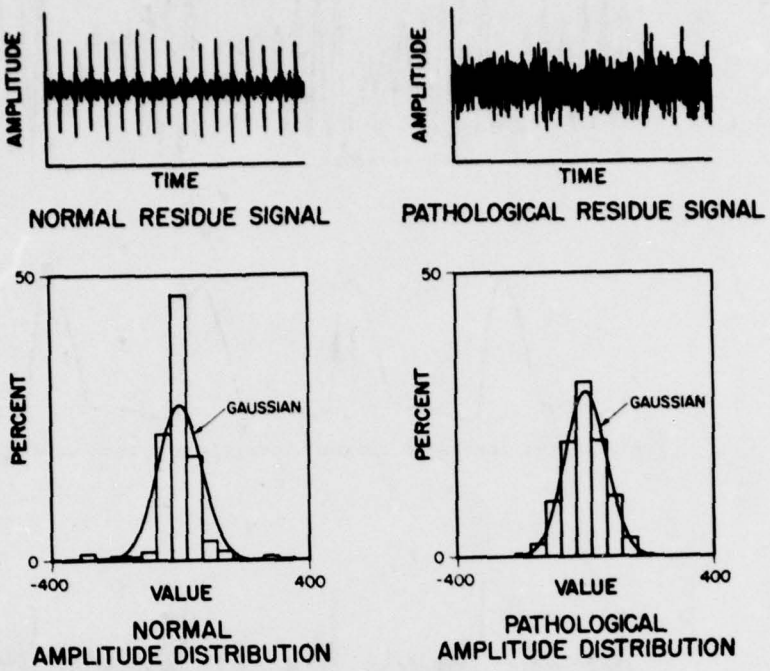


Figure 3: Amplitude distribution of normal and pathological residue signals (Davis, 1976b). The decreased noise in the normal residue signal results in an amplitude distribution that is narrow and taller than the pathological amplitude distribution.

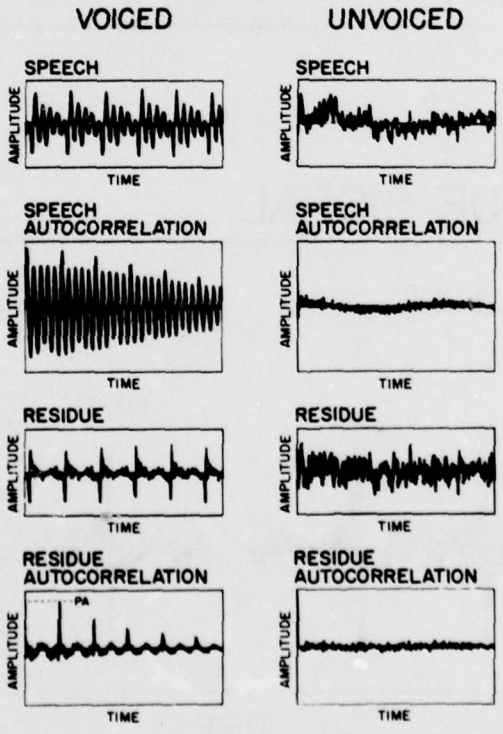


Figure 4: Comparison of voiced and unvoiced autocorrelation functions (Davis, 1976b). A pathological voiced sound may be like an unvoiced normal sound, and the PA will be low or nonexistent.

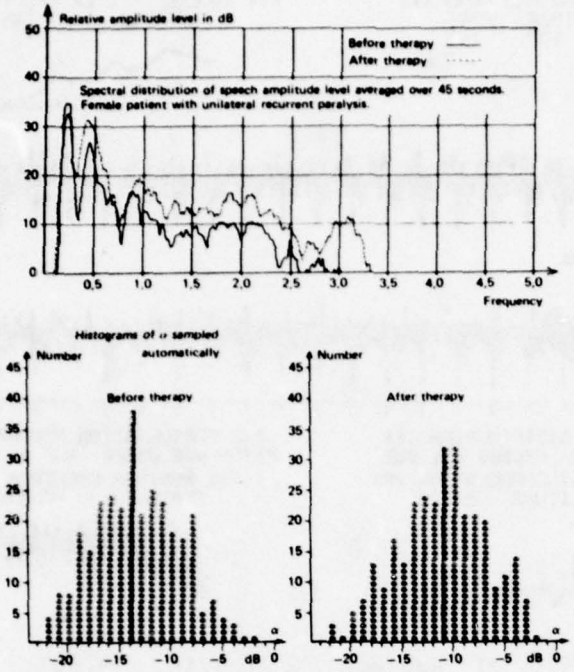


Figure 5: Comparison of long-term average spectra before and after voice therapy (Frøjær-Jensen and Prytz, 1976). There is a 3 dB increase in the energy above 1000 Hz relative to the energy below 1000 Hz after therapy.

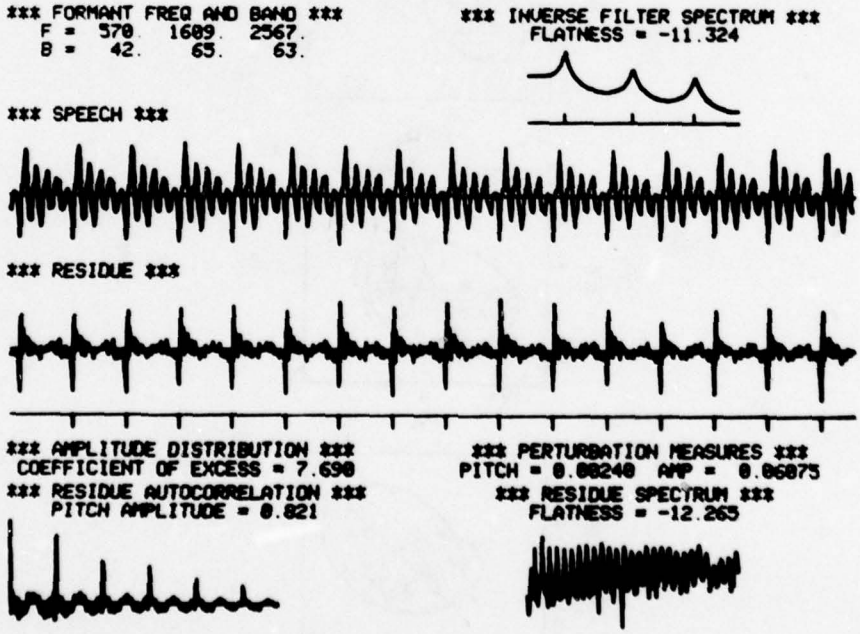


Figure 6: Voice Profile of subject N1 (Davis, 1976b). The waveforms and measures for this subject exemplify characteristics for a normal healthy voice.

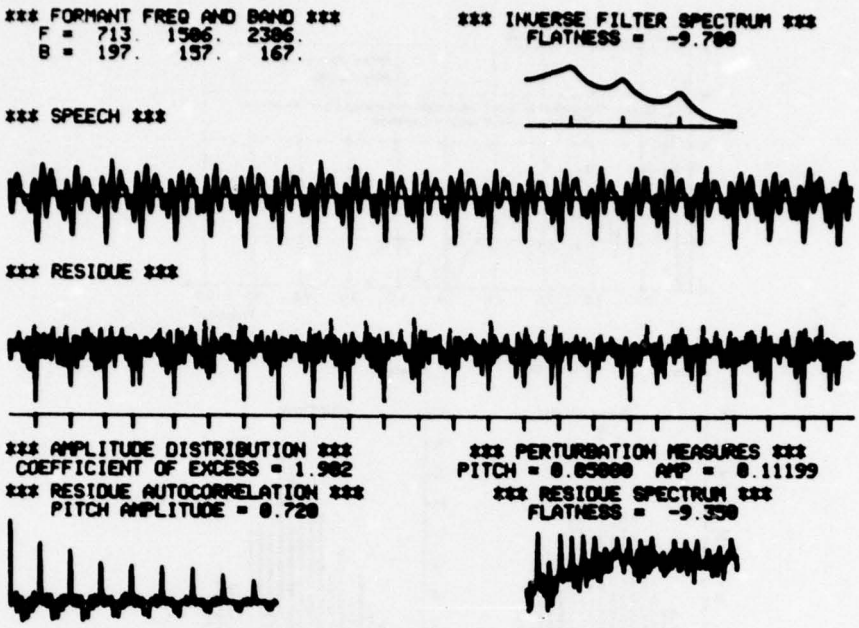


Figure 7: Voice Profile of subject P2 (Davis, 1976b). The subject had unilateral paralysis, and the pathology may have been undetected by auditory perception alone.

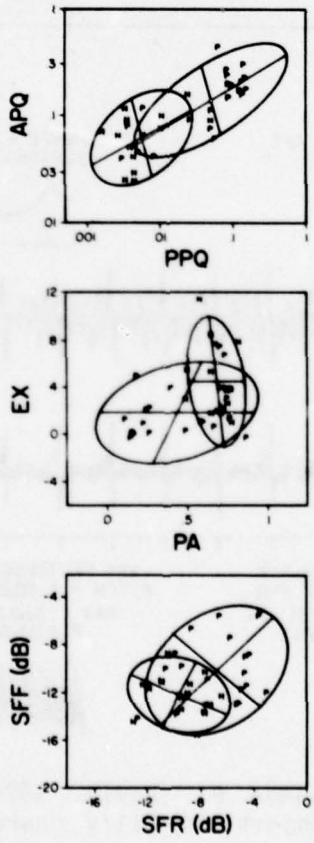


Figure 8: Scatter of normal and pathological features, showing two-sigma ellipses for each group (N = normal, P = pathological) (Davis, 1976b).

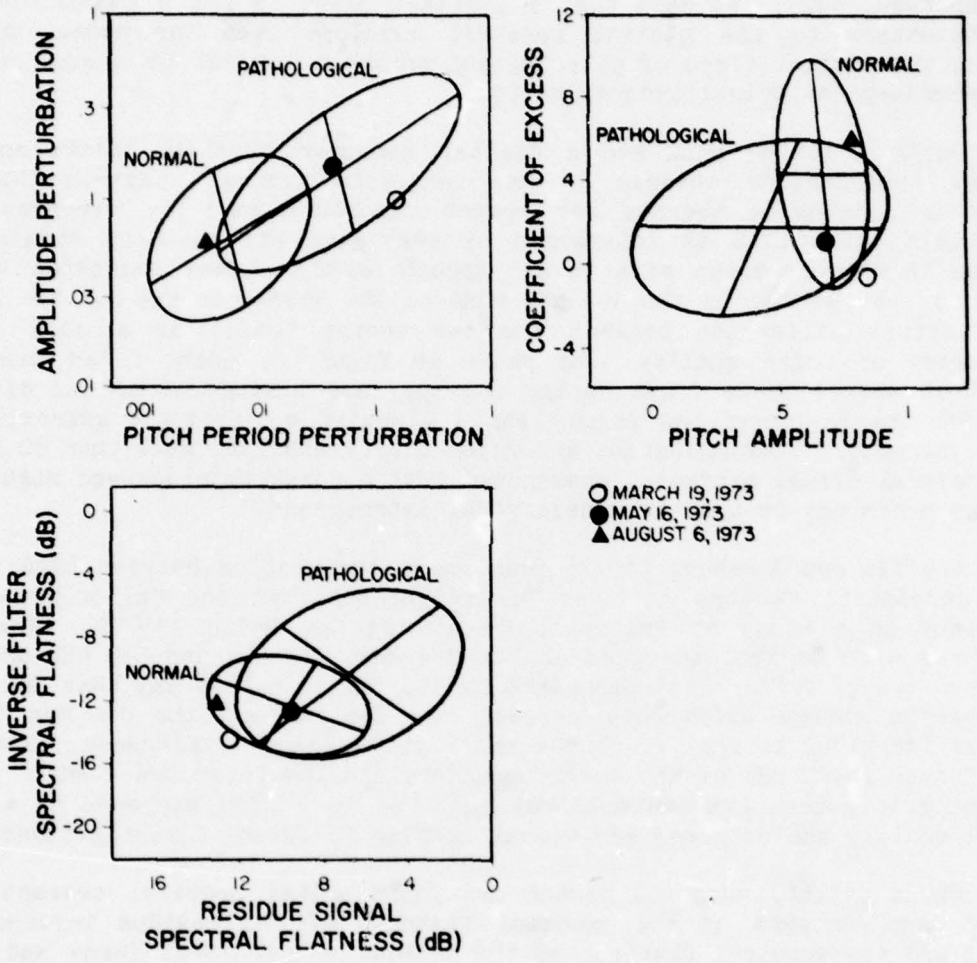


Figure 9: Comparison of features for voice therapy subject with normal and pathological ellipses (Davis, 1977, see footnote 4). The PPQ, APQ and EX show the best improvement.

studies of sound spectra indicate wide vocal variability, but they also illustrate the feasibility of using the acoustic spectrum to analyze laryngeal pathology.

In an attempt to quantify visual judgments of spectrograms, Hiki et al., (1976) measured the spectral slope of the glottal signal as a feature for acoustic analysis, but they did not present any quantitative results. Other investigators, however, using a "reflectionless" tube (Fisher, Monsen and Engebretson, 1975) indicate that a constant slope is not a particularly good approximation to the glottal spectral envelope even for normal speakers. Hence, the glottal slope of pathological speakers may not be a good parameter for normal-pathological discrimination.

Using a filter bank and a digital computer, Frøkjær-Jensen and Prytz (1976) investigated changes in the long-term average spectrum (LTAS) for patients undergoing therapy for speech disorders such as recurrent nerve paralysis. The LTAS is determined by averaging the spectrum obtained from voiced 80 msec segments of a 45 sec speech sample. They suggested that the ratio of the energy in the 1-5 kHz band to the energy in the 0-1 kHz band (or the decibel difference between the two energy bands) is a good spectral parameter of voice quality. As shown in Figure 5, there is an increase in spectral energy above 1 kHz during therapy, and histograms of the difference between the high and low energy bands (labeled α) indicate approximately a 4 dB increase. Frøkjær-Jensen and Prytz (1976) analyzed more than 50 patients and several normal subjects, and showed that a comparison between high and low energy bands may be used to assess vocal improvement.

Gauffin and Sundberg (1977) found some correlation between LTAS features and perceptual factors such as "overtight-breathy" and "hyper-hypokinetic" obtained in a study by Fritzell, Hammarberg and Wedin (1977). Their LTAS features were decibel energies in the 0-2 kHz, 2-5 kHz and 5-8 kHz bands, and decibel energy differences among the bands. It is noteworthy that Gauffin and Sundberg's energy difference between the 2-5 kHz and the 0-2 kHz bands is nearly identical to Frøkjær-Jensen and Prytz's energy difference, except that the former use 2 kHz as the energy boundary and the later use 1 kHz. The idea of using long-term average spectral measures is a good approach to assessing voice quality and requires additional testing in future investigations.

Davis (1976b) measured normal and pathological spectral characteristics using the concepts of the spectral flatness of the residue inverse filter (SFF) and the spectral flatness of the residue signal (SFR) (Gray and Markel, 1974). Spectral flatness is defined as the ratio in decibels of the geometric mean of the spectrum to the arithmetic mean of the spectrum. Gray and Markel observed that the more noiselike a spectrum, the greater its spectral flatness, having a maximum value of 0 dB for a constant spectrum. Unvoiced sounds, for example, fricatives, which are produced with an open glottis and a significant vocal tract constriction, have greater SFFs than voiced sounds, for example, steady vowels. Since the spectrum of the residue signal is essentially flat, showing only the fine spectral behavior of F_0 and its harmonic components, unvoiced sounds will have lower SFRs than voiced sounds. This result is a consequence of the harmonic nature of normal voiced speech; there are large negative excursions of the residue signal spectrum for each harmonic. As the sound becomes more noiselike (pathological), the harmonic

structure becomes less significant, and the SFR increases.

Using a linear model of speech production, it can be shown that the SFF is the negative sum of the spectral flatnesses of the lip radiation, vocal tract and glottal shaping spectra. If the vocal tract and lip radiation spectra are independent of laryngeal pathology, then changes in the glottal shaping spectrum caused by pathology will be measured by the SFF. It can also be shown that the SFR may be determined by subtracting the SFR from the spectral flatness of the speech spectrum (Gray and Markel, 1974).

Yanagihara (1967a, 1967b) noted the presence of noise components in pathological speech that mask formant characteristics and F_0 harmonics. Davis (1976b) used this observation as a basis for choosing the SFF and the SFR as spectral measures for vocal assessment. Davis assumed that the SFF is a measure of the masking of formant frequency amplitudes and bandwidths by noise, and that the SFR is a measure of the masking of F_0 harmonic amplitudes by noise. Since the vocal tract is assumed to be fixed and independent of the source excitation for a steady vowel, these masking effects may be attributed to changes in the sound source harmonic amplitudes caused by variations in the OQ and the amount of source turbulence. The relationship between the OQ and the harmonics of a periodic signal is evident from Fourier analysis; as the OQ increases, the amplitudes of the higher harmonics decrease. Physiologically, the OQ and the amount of turbulence are affected by any pathology interfering with the normal vibratory pattern of the vocal folds. Such effects may be caused by weak muscle action, or changes in the mass, elasticity or compliance of the folds.

VOICE PROFILES

Several studies mentioned above relate acoustic features to listener judgments of voice quality. The results of these studies demonstrate that acoustic parameters such as average pitch period perturbation are significantly correlated with perceptual parameters such as "hoarseness." However, at least one basic question remains. That is, can acoustic parameters alone provide measures that are clinically useful for evaluating pathological conditions in the larynx? For early detection or therapeutic assessment of the pathological voice, it is important to use such parameters in an easily-applied quantitative procedure. Davis (1975, 1976b) suggested that a profile of acoustic characteristics would be as useful to the laryngologist and speech pathologist as an audiogram is to the audiologist. Using features and signals obtained with residue inverse filtering, he developed a Voice Profile to display acoustic information about the voice and to serve as a record in a patient's medical history.

Davis[3] examined the usefulness of the Voice Profile by comparing the visual and numerical information conveyed in the Voice Profiles of normal and pathological subjects (Davis, 1976b) with qualitative observations of the acoustic characteristics of the same subjects (Koike and Markel, 1975). Koike

---

[3]Davis, S. B. (1978) Acoustic measures of laryngeal pathology using inverse filtered speech. Unpublished manuscript.

and Markel described representative residue signals selected from a data base of 10 normal and 10 pathological subjects, and although they indicated that the residue signal could be used to produce acoustic measures of laryngeal function, they did not make any measurements themselves. Davis (1976b) actually made the acoustic measurements and demonstrated that the measures could effectively discriminate between normal and pathological speakers. The data base used by Koike and Markel (1975) and Davis (1976b) is summarized in Table 1, and the acoustic features determined by Davis are listed in Table 2. The Voice Profiles for two of the subjects, N1 and P2, are shown in Figures 6 and 7, respectively. The Voice Profiles display six acoustic features (PPQ, APQ, EX, PA, SFF and SFR) and the signals required to compute the features. The following discussion is based on Davis' 1978 (see footnote 3) comparisons between Koike and Markel's (1975) descriptions and Davis' (1976b) measurements.

Normal Voice Sample

Figure 6 shows the Voice Profile for subject N1. Koike and Markel observed that the speech and residue signals are very regular, with each signal indicating a relatively small amplitude perturbation. They noted that this amplitude perturbation is greater than for subject N7 (not shown), and Davis measured an APQ of 6.08 percent for this subject, as compared to a measured APQ of 2.50 percent for subject N7. Koike and Markel commented that the residue signal has a high signal-to-noise ratio and near-constant periodicity. Davis found that the EX for this subject is 7.69, which is the third highest value in the normal group, and the PPQ for this subject is 0.24 percent, which is the lowest value in both groups. Thus, Davis' acoustic measurements for this normal subject correlate with Koike and Markel's qualitative descriptions.

Pathological Voice Sample

Figure 7 shows the Voice Profile for subject P2. Koike and Markel observed that this slightly hoarse subject represents a case of early laryngeal pathology that possibly would be undetected by auditory perception alone, but that might be detected with good acoustic measures of the residue signal. They indicated that the speech signal shows good periodicity and regularity among pitch periods, but the residue signal shows poor periodicity and a low signal-to-noise ratio. Davis found that the PPQ and APQ for subject P2 are higher than the corresponding measures for all but one of the normal subjects (N5), and the EX is lower than for all but two of the normal subjects (N5 and N10), thus again substantiating Koike and Markel's observations. Also, the observation that the pitch periods are visually similar to one another is confirmed by a high PA of 0.72, a value that is exceeded by only one other pathological subject (P6), and by only three of the normal subjects (N1, N9 and N10).

A comparison of the inverse filter spectra for subjects P2 and N1 shows noticeable differences (Figures 6 and 7). For subject N1, the peaks and valleys of the spectrum are distinct, the bandwidths are small, and the SFF is a low value of -11.3 dB. For subject P2, the valleys between the formant peaks are more shallow, the bandwidths are larger, and the SFF is a higher value of -9.8 dB. These broadband differences would probably be observed in

TABLE 1: Description of normal and pathological subjects age, sex, fundamental frequency and diagnosis.

Case	Age	Sex	F_0	Diagnosis
N1	29	M	110	Normal
N2	27	M	92	Normal
N3	30	M	120	Normal
N4	33	M	138	Normal
N5	26	M	140	Normal
N6	24	M	106	Normal
N7	31	M	124	Normal
N8	27	M	130	Normal
N9	23	F	232	Normal
N10	36	F	180	Normal
P1	33	F	205	Vocal nodule
P2	16	M	175	Unilateral paralysis
P3	56	M	96	Hemilaryngectomized
P4	25	F	196	Vocal nodule
P5	39	F	231	Spastic dysphonia
P6	39	M	115	Vocal polyp
P7	77	M	164	Laryngeal papilloma
P8	28	F	189	Unilateral paralysis
P9	57	M	---	Glottic cancer
P10	64	M	---	Advanced laryngeal cancer

TABLE 2: Acoustic features for normal and pathological subjects. The six principle features are the pitch period perturbation quotient (PPQ), amplitude perturbation quotient (APQ), pitch amplitude (PA), coefficient of excess (EX), spectral flatness of the inverse filter (SFF) and spectral flatness of the residue signal (SFR).

Case	PPQ(%)	APQ(%)	PA	EX	SFF(dB)	SFR(dB)
N1	0.24	6.08	0.82	7.69	-11.3	-12.3
N2	0.45	4.17	0.65	8.30	-10.2	-7.6
N3	0.47	5.42	0.71	3.92	-12.4	-9.4
N4	0.59	6.70	0.72	2.24	-11.3	-8.7
N5	0.34	11.63	0.65	1.74	-13.6	-7.8
N6	0.37	8.72	0.64	5.23	-13.4	-9.4
N7	0.46	2.50	0.58	11.33	-12.2	-6.3
N8	0.34	2.56	0.67	7.54	-8.6	-10.4
N9	0.51	4.59	0.77	3.77	-11.2	-11.9
N10	5.01#	9.04	0.77	0.96	-11.1	-10.8
P1	2.61	9.07	0.66	0.63	-9.9	-10.5
P2	5.08	11.20	0.72	1.90	-9.8	-9.3
P3	9.60	19.12	0.25	2.05	-8.8	-4.3
P4	1.87	15.33	0.60	0.97	-6.6	-8.6
P5	0.85	4.18	0.71	7.27	-5.6	-6.5
P6	0.60	11.68	0.74	2.98	-8.7	-9.8
P7	3.29	10.98	0.58	1.07	-13.0	-7.6
P8	13.25	14.71	0.17	-0.05	-7.4	-5.7
P9	10.68	16.00	0.49	0.28	-10.9	-7.4
P10	13.64	15.69	0.26	0.09	-5.3	-3.5

pitch period tracking errors

visual spectrogram analysis (Yanagihara, 1967a, 1967b; Gould, 1975).

A similar comparison applies between the residue spectra for subjects N1 and P2 (Figures 6 and 7). The harmonic nature of the voice source is readily apparent for the normal subject, and the SFR is a low value of -12.27 dB. In contrast, the residue spectrum for subject P10 exhibits an aperiodic harmonic structure (and hence a more noisy appearance), and the SFR is a higher value of -9.35 dB. These source harmonic differences would also probably be observed in visual spectrogram analysis.

Statistical Analysis of Normal and Pathological Data

In determining the advantages and limitations of these six acoustic features, Davis (1976b) used the data from the ten normal and ten pathological subjects and data from an additional seven normal and eleven pathological subjects in a statistical analysis. The additional subjects had characteristics similar to the first subjects, and were included to increase the population size so that statistical results would be significant. The means, standard deviations and correlations for the normal and pathological groups are listed in Tables 3 and 4.

A t-test (Bruning and Kintz, 1968) shows that the normal and pathological means of the PPQ, APQ and EX are significantly different at the 97.5 percent confidence level, the means of the PA and SFR are significantly different at the 95.0 percent level, and the means of the SFF are significantly different at the 90.0 percent level. Therefore, the PPQ, APQ and EX are the best features for distinguishing between these normal and pathological groups. Also, all of the differences between respective means have the correct sign, for example, the mean normal PPQ is less than the mean pathological PPQ, and the mean normal EX is greater than the mean pathological EX. However, the difference between the normal and pathological means of the SFF is insignificant.

Using Pearson's r score (Bruning and Kintz, 1968), the correlation matrices show two important relationships. For normal and pathological speakers, the correlation between the PPQ and APQ is positive (+0.826) and significant (at the 99.5 percent level). This correlation probably arises from the physical source of abnormal vocal fold vibrations, that is, a change in the mechanical properties of the affected tissues, since this change will cause both pitch period and amplitude perturbations. For normal and pathological speakers, the correlation between the PA and SFR is negative (-0.812) and significant (at the 99.5 percent level). This correlation may be explained as follows. A decrease in the PA indicates more noise and less periodicity in the residue signal (analogous to the generation of unvoiced fricatives), which indicates more noise and less harmonic structure in the residue spectrum, and consequently an increase in the SFR.

In Figure 8, the acoustic features for all subjects are cross-plotted by pairs together with two-sigma ellipses that are derived by a principal components analysis (Davis, 1976b). The axes of each ellipse intersect at the class means and represent orthogonal directions for the scatter of the data. The directions minimize the variance of the data within each normal or pathological class. The nonorthogonal appearance of the axes arises from the

TABLE 3: Statistics for pooled normal speakers.

	PPQ(%)	APQ(%)	PA	EX	SFR(dB)	SFF(dB)
Mean	0.99	7.22	0.725	5.17	-10.50	-11.849
S.D.	2.28	4.47	0.105	4.29	2.50	1.84

Correlation

	APQ	PA	EX	SFR	SFF
PPQ	0.826*	-0.115	-0.258	0.161	0.039
APQ		-0.100	-0.409	0.136	-0.277
PA			-0.015	-0.812*	0.164
EX				0.262	0.369
SFR					0.018

* = significant at the 99.5% confidence level

TABLE 4: Statistics for pooled pathological speakers.

	PPQ(%)	APQ(%)	PA	EX	SFR(dB)	SFF(dB)
Mean	4.54	11.99	0.599	2.44	-9.11	-11.851
S.D.	4.93	6.90	0.236	2.32	3.49	3.35

Correlation

	APQ	PA	EX	SFR	SFF
PPQ	0.625#	-0.615#	-0.564#	-0.491	0.064
APQ		-0.687*	-0.314	0.607#	0.070
PA			0.240	-0.869*	-0.098
EX				-0.011	0.060
SFR					0.137

* = significant at the 99.5% confidence level

= significant at the 99.0% confidence level

use of different scale factors for each axis. In the PPQ-APQ graph, logarithmic scaling is used since linear scaling does not adequately distinguish the normal and pathological classes. The use of logarithmic scaling is noteworthy because Kitajima et al. (1975) suggested that the use of a semitone scale (logarithmically-based) is a better basis for quantifying the auditory perception of F_0 perturbation. As a minor point, the mean PPQ and APQ in Figure 8 are different from their respective values in Tables 3 and 4 since the mean of the logarithm of the values is computed for the principal components analysis rather than the logarithm of the mean of the values.

In Figure 8, it is apparent that the normal and pathological classes are best distinguished by the perturbation quotients and least distinguished by the spectral flatness measures. However, even for the perturbation quotients, there are particular points (not indicated) that do not cluster in or near the correct group. Additionally, a normal speaker may have an abnormal value (outside a normal speaker ellipse) in one or possibly two dimensions, while measures in the remaining dimensions may be normal (inside a normal speaker ellipse). For this small number of subjects, it is unrealistic to expect all normal and pathological speakers to fall into tightly-clustered groups. Larger populations grouped by age and sex might yield more representative clusters of normal and pathological data. Also, these results indicate the multidimensional nature of the acoustic detection problem, with some features contributing more information than others for different normal and pathological speakers.

QUANTITATIVE ASSESSMENT OF VOICE THERAPY

The usefulness of acoustic features for the assessment of changes in voice quality can be determined by measuring changes in the features over time. In a pilot study, Davis[4] analyzed acoustic features for a patient undergoing voice therapy following removal of a vocal polyp. A trained listener subjectively observed that the voice quality continuously improved during the period of therapy. The data are summarized in Table 5 and compared with the distribution of the earlier data in Figure 9.

TABLE 5: Acoustic features for a voice therapy patient.

Date	PPQ(%)	APQ(%)	PA	EX	SFF(dB)	SFR(dB)
3/19/73	16.80	10.67	0.86	-0.30	-14.1	-12.7
6/16/73	4.85	15.51	0.71	1.48	-12.4	-9.6
8/ 6/73	0.40	5.97	0.79	6.20	-12.3	-13.0

[4]Davis, S. B. (1977) Acoustic analysis of voice pathology. American Speech and Hearing Association Annual Convention, Chicago.

The PPQ, APQ, and EX are the features in the t-test that show the most significant separation between the normal and pathological mean values. These features also show the best improvement during therapy for this patient. In comparison with the normal ellipses in Figure 9, these parameter values shift from lying outside the normal range at the beginning of therapy to lying within the normal range after several months of therapy. Thus, the results indicate that changes in these features correlate with changes in voice quality for this patient.

None of the other features show changes that correlate with voice quality improvement. For the earliest session, even though the voice quality is poor, and the PPQ is high, the PA is higher than that of any normal or pathological subject listed in Table 2. Visual observation of the speech signal reveals that there is a very high degree of regularity between adjacent pitch periods. The residue signal has a low EX and a very noisy appearance, but the high degree of pitch period regularity observed in the speech signal is maintained in the residue signal and leads to a high PA. Thus the PA may not be as good a measure as the PPQ, APQ or EX as a measure of improvement during therapy.

The SFF and SFR for this subject show no trends that can be correlated with the data used to derive the ellipses. The SFF decreases, and the SFR fluctuates during therapy for this patient. Such trends suggest that these features may be inappropriate for quantifying observations from spectrograms, and perhaps modified or new features would reflect spectral changes more accurately.

Thus, the PA, SFF and SFR show no consistent improvement during therapy and their normal and pathological mean values show less significant separation, but future testing is needed before rejecting them or reducing their weight in an overall assessment of voice quality.

SUMMARY AND FUTURE INVESTIGATIONS

The results discussed in this chapter indicate the feasibility of using quantifiable acoustic features to distinguish between normal and pathological subjects. The acoustic features relating to laryngeal function provide more information when inverse filtering is used to remove the supraglottal structure from the speech signal. These features can be automatically computed from a digitized representation of the signal, and the results may be organized to form a Voice Profile.

However, the application of digital analysis techniques poses some difficulties. A linear model of speech production is used to derive the inverse-filtered signal, but the assumption of minimum glottal source-vocal tract interaction may be tenuous for some pathological conditions and requires further study. Nevertheless, the assumption of independence is reasonable for normal and mildly pathological subjects, and these subjects are the ones for whom acoustic analysis is potentially a valuable addition to existing medical procedures.

A further problem is that different sustained vowel sounds will result in different perturbation values (Johnson[5]). This effect is probably a consequence of the interaction between the glottal source and vocal tract, and indicates the continuing need for study of the complex relationships between the glottal sound source and the acoustic characteristics of the vocal tract. Since any single study uses the same vowel for all speakers, the effect of vowel type on the acoustic measures is probably uniformly distributed among the speakers, and therefore not a significant source of error. Further analysis is also required to determine how consistently acoustic features may be measured from independent samples of the same vowel from the same speaker. Additional statistics should be collected for both sexes and among different age groups.

An additional goal is to analyze acoustic features in a clinic and to detect early cases of laryngeal pathology. Voice Profiles may be useful indicators of voice quality, especially during voice therapy, and it will be necessary to have speech pathologists and physicians evaluate their usefulness. For efficient clinical implementation, tape-recorded voice samples may be sent from a clinic to a central computer facility, either indirectly via the mail system or directly via telephone lines. With the advent of hospital computers and remote terminals in outpatient clinics, an immediate acoustic evaluation of laryngeal pathology is a viable objective. Alternatively, voice samples could be analyzed with a microprocessor-based "black box" built especially for clinical use.

Further study should be directed to at least two problems: the identification of additional features useful in acoustic evaluation of laryngeal pathology, and the possibility of discriminating among types and degrees of pathology. Additional measures might involve the use of long-duration or short-duration voice samples and might include, for example: a) formant frequencies, bandwidths and amplitudes; b) the amplitude of the first harmonic of F_0 in the residue signal autocorrelation; c) the slope of the line through the peaks in the residue signal autocorrelation; d) the periodicity and peak amplitudes of the pitch period and amplitude correlograms. Some of these features have been suggested previously (Hiki et al., 1976), but neither acoustic or physiological significance nor clinical analysis methods have been established yet for any of these additional features.

Finally, acoustic features should be compared with subjective voice quality ratings (Murry, 1975). Both real and synthesized pathological speech samples (with known acoustic deviations) could be examined. Also, other parameters, for example, the Euclidean distance between a given feature vector and an "ideal" feature vector (having zero perturbation measures, unity pitch amplitude, etc.), might be analyzed as measures of voice quality.

---

[5]Johnson, K. W. (1969) The effect of selected vowels on laryngeal jitter. Master's thesis, University of Kansas, Lawrence.

It is evident that acoustic voice analysis using inverse filtering may be used for screening individuals for the early detection of laryngeal pathology or for assessing improvement during voice therapy. The fields of speech pathology and laryngology will benefit significantly from future research on the acoustic characteristics of normal and pathological voices.

REFERENCES

Arnold, G. (1955) Vocal rehabilitation of paralytic dysphonia: II. Acoustic analysis of vocal function. Archives of Otolaryngology 62, 593-601.

Atal, B. S. and S. L. Hanauer. (1971) Speech analysis and synthesis by linear prediction of the speech wave. Journal of the Acoustical Society of America 50, 637-655.

van den Berg, Jw. (1962) Modern research in experimental phoniatrics. Folia Phoniatrica 14, 81-149.

Blackman, R. B. and J. W. Tukey. (1958) The Measurement of Power Spectra. (New York: Dover).

Bowler, N. W. (1964) A fundamental frequency analysis of harsh vocal quality. Speech Monographs 31, 128-134.

Bruning, J. L. and B. L. Kintz. (1968) Computational Handbook of Statistics. (Glenview, Illinois: Scott, Foresman and Company).

Chiba, T. and M. Kajiyama. (1941) The Vowel, Its Nature and Structure. (Tokyo: Tokyo Kaiseikan).

Cramer, H. (1958) Mathematical Methods of Statistics. (Princeton: Princeton University Press).

Crystal, T. H. and C. L. Jackson. (1970) Extracting and processing vocal pitch for laryngeal disorder detection. Journal of the Acoustical Society of America 48(A), 118.

Davis, S. B. (1975) Preliminary results using inverse filtering of speech for automatic evaluation of laryngeal pathology. Journal of the Acoustical Society of America 58(A), S111.

Davis, S. B. (1976a) Determination of glottal area based on digital image processing of high-speed motion pictures of the vocal folds. Journal of the Acoustical Society of America 60(A), S65.

Davis, S. B. (1976b) Computer evaluation of laryngeal pathology based on inverse filtering of speech. Ph.D. Dissertation, University of California, Santa Barbara. Also SCRL Monograph 13, Speech Communications Research Laboratory, Inc., Santa Barbara.

Fant, C. G. M. (1959) Acoustic Analysis and Synthesis of Speech with Applications to Swedish. Ericsson Technics 15, 3-108.

Fant, C. G. M. (1960) Acoustic Theory of Speech Production. ('s Gravenhage: Mouton and Co.).

Farnsworth, D. W. (1940) High Speed Motion Pictures of Human Vocal Cords. Record 18, Bell Laboratories, 203-208.

Fisher, W. M., R. B. Monsen and A. M. Engebretson. (1975) Variations in the normal male glottal wave. Journal of the Acoustical Society of America 58(A), S41.

Flanagan, J. L. (1958) Some properties of the glottal sound source. Journal of Speech and Hearing Research 1, 99-116.

Flanagan, J. L. (1972) Speech Analysis, Synthesis and Perception. (New York: Springer-Verlag).

Frank, D. I. (1940) Hoarseness - A new classification and a brief report of

four interesting cases. Laryngoscope 50, 472-478.
Fritzell, B., B. Hammarberg and L. Wedin. (1977) Clinical applications of acoustic voice analysis, Part I. Quarterly Progress and Status Report. (Stockholm: Speech Transmission Laboratory, Royal Institute of Technology), no. 2-3, 31-38.
Frøjær-Jensen, B. and S. Prytz. (1976) Registration of voice quality. Bruel and Kjær Technical Review no. 3, 3-17.
Gauffin, J. and J. Sundberg. (1977) Clinical applications of acoustic voice analysis, Part II. Quarterly Progress and Status Report. (Stockholm: Speech Transmission Laboratory, Royal Institute of Technology), no. 2-3, 39-43.
Gould, W. J. (1973) The Gould laryngoscope. Transactions of the American Academy of Ophthalmology and Otolaryngology 77, ORL-139-141.
Gould, W. J. (1975) Quantitative assessment of voice function in microlaryngology. Folia Phoniatrica 27, 157-165.
Gray, A. H. Jr. and J. D. Markel. (1974) A spectral flatness measure for studying the autocorrelation method of linear prediction of speech analysis. IEEE Transactions on Acoustics, Speech and Signal Processing ASSP-22, 207-217.
Hayden, E. and Y. Koike. (1972) A data processing scheme for frame by frame film analysis. Folia Phoniatrica 24, 169-186.
Hecker, M. H. L. and E. J. Kreul. (1971) Descriptions of the speech of patients with cancer of the vocal folds, Part 1: Measures of fundamental frequency. Journal of the Acoustical Society of America 49, 1275-1282.
Hiki, S., S. Imaizumi, M. Hirano, H. Matsushita and Y. Kakita. (1976) Acoustical analysis for voice disorders. Conference Record, 1976 IEEE International Conference on Acoustics, Speech, and Signal Processing. (Rome, N.Y.: Canterbury Press).
Hiki, S., K. Sugawara and J. Oizumi. (1968) On the rapid fluctuation of voiced pitch. Reports of the Research Institute of Electrical Communication 19. (Tohoku University, Japan), 237-239.
Holmes, J. N. (1962) An investigation of the volume velocity waveform at the larynx during speech by means of an inverse filter. Proceedings of the Speech Communication Seminar. (Stockholm: Speech Transmission Laboratory, Royal Institute of Technology), B-4.
Holmes, J. N. (1975) Low frequency phase distortion of speech recordings. Journal of the Acoustical Society of America 58, 747-749.
Isshiki, N. (1966) A method of classified description of hoarse voice, (in Japanese). Japanese Journal of Logopedics and Phoniatrics 7, 15-21.
Isshiki, N., H. Okamura, M. Tanabe and M. Morimoto. (1969) Differential diagnosis of hoarseness. Folia Phoniatrica 21, 9-19.
Jackson, C. and C. L. Jackson. (1937) The Larynx and its Diseases. (Philadelphia: W. B. Saunders).
(New York: John Wiley and Sons).
Kelly, J. L. and C. Lochbaum. (1962) Speech synthesis. Proceedings of the 4th International Congress in Acoustics G42, 1-4.
Kitajima, K., M. Tanabe and N. Isshiki. (1975) Pitch perturbation in normal and pathologic voice. Studia Phonologica IX, 25-32.
Koike, Y. (1967) Application of some acoustic measures for the evaluation of laryngeal dysfunction. Journal of the Acoustical Society of America 42(A), 1209.
Koike, Y. (1969) Vowel amplitude modulations in patients with laryngeal

diseases. Journal of the Acoustical Society of America 45, 839-844.
Koike, Y. (1973) Application of some acoustic measures for the evaluation of laryngeal dysfunction. Studia Phonologica 7, 17-23.
Koike, Y. and M. Hirano. (1973) Glottal area time function and subglottal pressure variation. Journal of the Acoustical Society of America 54, 1618-1627.
Koike, Y. and J. D. Markel. (1975) Application of inverse filtering for detecting laryngeal pathology. Annals of Otology, Rhinology and Laryngology 84, 117-124.
Koike, Y. and H. Takahashi. (1971) Glottal parameters and some acoustic measures in patients with laryngeal pathology. Studia Phonologica 7, 45-50.
Kolmogoroff, A. N. (1941) Interpolation und Extrapolation von stationären zufälligen Folgen. Bulletin de l'academie des sciences de U.S.S.R., Ser. Math 5, 3-14.
von Leden, H., P. Moore and R. Timcke. (1960) Laryngeal vibrations: Measurements of the glottic wave, Part III. The pathologic larynx. Archives of Otolaryngology 71, 16-35.
Levinson, N. (1946) The Wiener RMS (Root Mean Square) error criterion in filter design and prediction. Journal of Mathematical Physics 25, 261-278.
Lieberman, P. (1961) Perturbations in vocal pitch. Journal of the Acoustical Society of America 33, 597-603.
Lieberman, P. (1963) Some acoustic measures of the fundamental periodicity of normal and pathologic larynges. Journal of the Acoustical Society of America 35, 344-353.
Lindqvist, J. (1965) Studies of the voice source by means of inverse filtering technique. Proceedings of the 5th International Congress on Acoustics A35, Liege, September.
Luchsinger, R. and G. E. Arnold. (1965) Voice-Speech-Language. Clinical Communicology: Its Physiology and Pathology. (Belmont: Wadsworth).
Makhoul, J. I. (1975) Spectral linear prediction: Properties and applications. IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP-23, 283.
Makhoul, J. I. and J. J. Wolf. (1972) Linear Prediction and the Spectral Analysis of Speech, BBN Report no. 2304. (Cambridge: Bolt, Beranek and Newman, Inc.).
Markel, J. D. (1971) Formant Trajectory Estimation from a Linear Least-Squares Inverse Filter Formulation, SCRL Monograph no. 7. (Santa Barbara: Speech Communications Research Laboratory, Inc.).
Markel, J. D. (1972) Digital inverse filtering - A new tool for formant trajectory estimation. IEEE Transactions on Audio and Electroacoustics AU-20, 129-137.
Markel, J. D. (1973) Application of a digital inverse filter for automatic formant and F_0 analysis. IEEE Transactions on Audio and Electroacoustics AU-21, 154-160.
Markel, J. D. and A. H. Gray, Jr. (1973) On autocorrelation equations as applied to speech analysis. IEEE Transactions on Audio and Electroacoustics AU-20, 69-79.
Markel, J. D. and A. H. Gray, Jr. (1976) Linear Prediction of Speech. (New York: Springer Verlag).
Markel, J. D., A. H. Gray, Jr. and H. Wakita. (1973) Linear Prediction of

Speech - Theory and Practice, SCRL Monograph no. 10. (Santa Barbara: Speech Communications Research Laboratory, Inc.).
Miller, R. L. (1959) Nature of the vocal cord wave. Journal of the Acoustical Society of America 31, 667-677.
Moore, G. P. (1938) Motion picture studies of the vocal folds and vocal attack. Journal of Speech and Hearing Disorders 3, 235-238.
Moore, G. P. (1968) Otolaryngology and speech pathology. Laryngoscope 78, 1500-1507.
Moore, G. P. (1971) Voice disorders organically based. In Handbook of Speech Pathology and Audiology, ed. by L. Travis. (New York: Appleton Century Croft).
Moore, G. P. (1973) Terminal Report for a Conference on Early Detection of Laryngeal Pathology. (Gainesville: Communications Sciences Laboratory, Department of Speech, University of Florida).
Murry, T. (1975) Some acoustic features of hoarseness. Journal of the Acoustical Society of America, 58(A), S111.
Nessel, E. (1960) Über das Tonfrequenzspectrum der pathologisch veränderten Stimme. Acta Oto-Laryngologica (S)157.
Osgood, C. E., G. J. Suci and P. H. Tannenbaum. (1957) The Measurement of Meaning. (Urbana: University of Illinois Press).
Palmer, J. M. (1959) Hoarseness in laryngeal pathology, a review of the literature. Laryngoscope 61, 500-516.
Perkins, W. H. (1971) Vocal function: A behavioral analysis. In Handbook of Speech Pathology and Audiology, ed. by L. Travis. (New York: Appleton Century Croft).
Peterson, G. E. and H. L. Barney. (1952) Control methods used in a study of vowels. Journal of the Acoustical Society of America 24, 175-184.
Ramsey, J. L. (1964) Logical Techniques for Glottal Source Measurements. Instrumentation Papers no. 48, Data Sciences Laboratory Project 5628. (Hanscom Field, Mass.: Air Force Cambridge Research Laboratories).
Rontal, E. (1975) 'Picturing' vocal cord therapy results. Journal of the American Medical Association 233, 1149-1150.
Rothenberg, M. (1973) A new inverse filtering technique for deriving the glottal air flow waveform during voicing. Journal of the Acoustical Society of America 53, 1632-1645.
Sawashima, M. and H. Hirose. (1968) New laryngoscopic technique by use of fibre optics. Journal of the Acoustical Society of America 43, 168-169.
Schönhärl, E. (1960) Die Stroboskopie in der praktischen Laryngologie. (Stuttgart: George Thieme).
Smith, W. and P. Lieberman. (1964) Studies in Pathologic Speech Production, Final Report, AFCRL-64-379. (Hanscom Field, Mass.: Air Force Cambridge Research Laboratories).
Sondhi, M. M. (1975) Measurement of the glottal waveform. Journal of the Acoustical Society of America 57, 228-232.
Soron, H. I. (1967) High-speed photography in speech research. Journal of Speech and Hearing Research 10, 768-776.
Takahashi, H. and Y. Koike. (1975) Some perceptual dimensions and acoustical correlates of pathologic voices. Acta Oto-Laryngologica S338, 1-24.
Takasugi, T. and J. Suzuki. (1970) Consideration of voice source, In "Analysis by Synthesis Technique." Journal of the Radio Research Laboratory 17, 153-168.
Tanabe, M., K. Kitajima, W. J. Gould and A. Lambiase. (1975) Analysis of high

speed motion pictures of the vocal folds. Folia Phoniatrica 27, 77-87.
Timcke, R., H. von Leden and P. Moore. (1958) Laryngeal vibrations: Measurements of the glottic wave. Part I. The normal vibratory cycle. AMA Archives of Otolaryngology 68, 1-19.
Timcke, R., H. von Leden and P. Moore. (1959) Laryngeal vibrations: Measurements of the glottic wave. Part II, Physiologic variations. AMA Archives of Otolaryngology 69, 438-444.
Weiner, N. (1949) Extrapolation, Interpolation and Smoothing of Stationary Time Series. (Cambridge: M.I.T. Press).
Wendahl, R. W. (1963) Laryngeal analog synthesis of harsh voice quality. Folia Phoniatrica 15, 241-250.
Wendahl, R. W. (1966) Laryngeal analog synthesis of jitter and shimmer, auditory parameter of harshness. Folia Phoniatrica 18, 98-108.
Winckel, F. (1952) Electroakustische Untersuchungen an der menschlichen Stimme. Folia Phoniatrica 4, 93-113.
Winckel, F. (1954) Physikalische Kriterien fur objektive Stimmbeurteilung. Folia Phoniatrica 5, 232-252.
Wong, D. W. and J. D. Markel. (1976) Considerations in the estimation of glottal volume velocity waveforms. Journal of the Acoustical Society of America 59(A), S96-S97.
Yanagihara, N. (1967a) Significance of harmonic changes and noise components in hoarseness. Journal of Speech and Hearing Research 10, 531-541.
Yanagihara, N. (1967b) Hoarseness: Investigation of the physiological mechanisms. Annals of Otology 76, 472-489.
Zemlin, W. R. (1968) Speech and Hearing Science. (Englewood Cliffs: Prentice Hall).

Speech Synthesis by Rule Using the FOVE Program*

Frances Ingemann+

ABSTRACT

The FOVE program for synthesizing English is briefly described and the results of various intelligibility tests are presented. In meaningful sentences, word intelligibility of 84-85 percent and phoneme intelligibility of 89-91 percent were achieved. In syntactically normal nonsense sentences, word intelligibility of 77-82 percent and phoneme intelligibility of 89-91 percent were achieved. However, in isolated words from sets with minimal differences, only 86 percent phoneme intellibility was achieved.

A program for speech synthesis that uses the OVE III synthesizer (Liljencrants, 1968) to generate the acoustic output signal has been developed at Haskins Laboratories. This program, called FOVE, has been designed to allow the user who is not a computer programmer to test various acoustic-phonetic rules for speech synthesis. FOVE is based on the OVEBORD program designed by Kuhn (1973), which in its turn was developed from a modified Holmes-Mattingly algorithm (Holmes, Mattingly and Shearme, 1964; Mattingly, 1968). Because speech can be synthesized without delay as soon as a set of rules is available and a phonetic text has been typed in, a researcher can work on the rules interactively by making a minor modification, listening to the output to determine the effect of the change, and deciding whether to retain the modification or to make other changes.

Table 1 shows the components used to synthesize speech. The input for a text consists of phonetic symbols for English phonemes, two stress marks and three kinds of intonation marks. Acoustic values for the phonetic symbols are assigned by a set of tables (one or more for each phonetic symbol), allophone rules, and invariant internal rules not accessible to the user for processing these user-specified variables.

Each table provides values for all the parameters on OVE III except the fundamental frequency. OVE III parameters with their ranges and steps are

---

*This paper was presented at the IPS-77, Miami Beach, and will appear in the congress proceedings.

+Also University of Kansas.

Acknowledgment: This work was supported by Veterans Administration contract V101 (134) P-342 and BRSG Grant RR-05596. I want to thank Patrick Nye for his suggestions in the preparation of this paper.

[HASKINS LABORATORIES: Status Report on Speech Research SR-54 (1978)]

TABLE 1: FOVE program for synthesizing speech using an OVE III synthesizer.

I. Input: phonemic transcription consisting of English phonemes plus two stress marks and three types of intonation marks.

II. User-specified values.

A. Phoneme tables

1. Name
2. Second name
3. Distinctive feature number
4. Duration (in time frames, usually 5 msec. each)
5. Initial voicing
6. Final hiss
7. Final friction
8. Hiss for initial portion of following segment

Values for OVE III Control Parameters

9. Amplitude of hiss
10. Amplitude of formant buzz
11. F_1 frequency
12. F_2 frequency
13. F_3 frequency
14. F_1 bandwidth
15. F_2 bandwidth
16. F_3 bandwidth
17. Amplitude of nasal buzz
18. Frequency of nasal buzz
19. Amplitude of friction
20. Pole/zero ratio
21. Fricative pole 1 frequency
22. Fricative pole 2 frequency

Weighting of adjacent values to determine boundary values

23-36. Weight of present phoneme values (9-22)
37-50. Weight of adjacent phoneme values (9-22)

Transition durations

51-64. Duration of transitions for parameters 9-22 in the present phoneme
65-78. Duration of transitions for parameters 9-22 in the adjacent phoneme

B. Allophone rules changing any phoneme table value in a specified context.

C. Fundamental frequency values associated with stress and intonation marks.

III. Determination of parameter values for each time frame from

A. User specified values

B. Internal rules that include

1. Dominance of one phoneme over another
2. Syllabification
3. Location of tonic syllable and resultant pitch assignments.

listed in Table 2. The phoneme table also provides values for durations (numbers 4-8 in Table 1) and transitions (numbers 23-78). In addition, the phoneme table assigns each phoneme a so-called distinctive feature number, whereby sounds can be grouped into traditional phonetic classifications for use in allophone rules and to determine which of two adjacent phonemes will provide weighting and durational values for transitions. Mattingly (1968) has described the assignment and use of these distinctive feature numbers in the program from which OVEBORD and FOVE are both derived.

TABLE 2: Summary of OVE III synthesis parameters.

Parameter	Range	Step
Fundamental frequency	50-315 Hz	.8%
Amplitude of hiss informants	0-30 dB	2 dB
Amplitude of periodic source	0-31.5 dB	.5 dB
First formant	200-1260 Hz	.8%
Second formant	500-3200 Hz	.8%
Third formant	1000-6300 Hz	.8%
F_1 bandwidth increment	65-188 Hz	12 Hz
F_2 bandwidth increment	66-470 Hz	31 Hz
F_3 bandwidth increment	76-750 Hz	250 Hz
Nasal amplitude	0-30 dB	2 dB
Nasal formant	200-1230 Hz	3%
Fricative amplitude	0-30 dB	2 dB
Pole/zero ratio	0-31.5 dB	.5 dB
Fricative formant 1	1000-6200 Hz	3%
Fricative formant 2	2500-15700 Hz	3%

Allophone rules allow any of the values of the phoneme tables to be changed in specified environments. The set of environments available is basically the same as that given in Mattingly (1968).

Prosodic information is provided by the stress and intonation markers. Durational differences are specified by allophone rules. Fundamental frequencies are determined by twelve pitch specifications. Three of these provide an initial frequency and the range within which the frequency may vary. The remaining nine are associated in a complex way with the major stress mark (the minor stress mark has no effect on fundamental frequency) and intonation markers. Six of the pitch specifications determine the steps by which the frequency may rise or fall in relation to the presence or absence of major stress marks. The remaining three specify the final frequency contour

starting at the last major stress.

There are, in addition, rules which the user may not vary that do such things as determine syllable boundaries, location of tonic syllable, and dominance of one phoneme table over another for purposes of deciding transition values.

To see how well speech could be synthesized using this program and to determine which values are most suitable for the phoneme tables and which allophone rules are needed, I have been working on a set of rules. Periodically, the rules have been tested on listeners unfamiliar with synthetic speech to find out what progress was being made and to see what specific sounds needed improvement. The overall results of these tests are given in Table 3. An example of the kind of speech synthesized by rule in early 1977 is a passage that contains the instructions for the Carnegie-Mellon University (CMU) sentences whose intelligibility scores are reported at the top of Table 3.

The CMU sentences are based (with minor modifications to split very long sentences into two shorter ones) on those devised by Shockey (1974) at Carnegie-Mellon University. They contain all the phonemes of English in at least two environments. The sentences are meaningful but frequently odd, so that correct responses are more dependent upon correct phoneme identification than they might be in more predictable sentences. Listeners unfamiliar with synthetic speech were asked to write down the sentences after hearing each sentence spoken twice. By 1977, listeners to FOVE-generated synthetic speech were able to get 84 percent of the words correct and, when intended phonemes were matched with responses, 91 percent of the phonemes correct, with vowels scoring slightly better than consonants.

Somewhat lower scores were obtained on the Mitchell test (Mitchell, 1974) for words in isolation that were heard only once. The Mitchell test consists of 4 lists of 50 monosyllabic words each, designed to test 22 consonants in initial position, 13 consonants in final position and 15 vowels and diphthongs. The listener is presented with a choice of five words differing only in one or two distinctive features. Because the test was originally devised to be used with natural speech to assess hearing impairments or transmission systems, the choices did not always make provision for the kind of errors listeners make with synthetic speech. For this reason, listeners were allowed to write down another answer if none of the words on the answer sheet corresponded to what they heard. Only the specific sound being tested in each word was scored. Under these conditions, phonemes were only 86 percent correct in contrast to 91 percent in sentences.

Because the CMU sentences contained very few examples of some sounds and others were omitted entirely from the Mitchell lists, for the 1976 rules a new set of 24 phonetically diverse (PD) sentences was devised to contain additional tokens of sounds poorly represented in the CMU sentences, and new contexts for all sounds. Although not so intended, these sentences turned out to be considerably easier as a set than the CMU sentences, scoring 88 percent word intelligibility with 92 percent phonemes correct. When scores are that high, it is difficult to detect improvement and so the set as a whole was not used for further testing. However, a subset of the PD sentences was formed by

TABLE 3: Percent correct for three sets of rules on various intelligibility tests.

	1974 Rules	1976 Rules	1977 Rules
CMU Sentences			
Words	75%	81%	84%
Phonemes	83%	86%	91%
Consonants	81%	86%	90%
Vowels	85%	88%	93%
Mitchell Lists			
22 Initial cons.	78%	79%	79%
13 Final cons.	77%	79%	84%
15 Vowels and diphthongs	97%	98%	99%
Total	84%	84%	86%
PD Sentences			
Words		88%	
Phonemes		92%	
SPD Sentences			
Words			85%
Phonemes			89%
Syntactically Normal Nonsense Sentences			
Words			77, 82%
Phonemes			89, 91%

TABLE 4: Average time spent answering questions and number of errors in a listening comprehension test.

	Time	Errors
1974 Rules (n = 6)	4.30 min.	2.0
Natural speech (n = 24)	4.52 min.	2.29

eliminating all but one of the sentences that had had fewer than two errors in the 1976 test. In addition, a few long sentences were shortened to delete clauses or phrases on which no errors had been made.[1] The number of sentences constituting the new subset of PD sentences (labeled SPD in Table 3) was thirteen. These sentences scored about the same on the 1977 rules as did the CMU sentences, with a slightly higher word correct score and a slightly lower phoneme score.

The last set of sentences reported in Table 3 are 50 syntactically normal nonsense sentences taken from Nye and Gaitenby (1974). These sentences consist of four monosyllabic English words in the frame 'The _____ _____ _____ the _____': for example, "The short arm sent the cow." These sentences reduce the ability to predict the test words on the basis of content. It should be noted, however, that a semantic effect could not be avoided entirely: a number of listener errors were in the direction of making the sentences more meaningful than the intended sentence. In this test the listeners were presented with the sentence frame on the answer sheet and asked to fill in the four blanks after hearing each sentence only once.

Two scores are given for these sentences. The first scores were obtained when the listeners heard these sentences at the beginning of the test session. The second scores are for the same sentences when they were preceded by 10 similar sentences used to test natural versus rule durations (Ingemann, 1977). Comparison of the two scores reveals an obvious learning effect. A second point to be noted is that although the word scores are somewhat lower than the meaningful sentences (77 percent and 82 percent in contrast to 84 percent for CMU sentences and 85 percent for SPD sentences), the phonemes were correctly identified at approximately the same rate in all three sets of sentences.

Another test involved extended listening to find out how well listeners could understand and obtain information from synthetic speech. A test comparing the 1974 rules with other synthesis versions and natural speech has been reported in Nye, Ingemann and Donald (1975). Only the results for 1974 rules, which scored the highest of the synthesis versions tested, will be summarized here.

Two texts from a published reading test for college-bound and college students were synthesized. One text was approximately 2000 words in length and the other 1700. Natural speech recordings of a male speaker were used as a control. Each listener heard one text synthesized and the other text in natural speech. After hearing a text played through once without interruption, the listeners were required to answer fourteen multiple choice questions. Listeners were allowed to selectively replay passages and check off answers until they were confident that all the questions had been answered correctly. The time the listeners took to answer the questions (including the replay time) was recorded. The results in Table 4 show that there is little difference between natural speech and the 1974 rules and the difference is not

---

[1]Four additional sentences are not tabulated here because they were subject to durational manipulations described in a paper reported elsewhere (Ingemann, 1977).

significant. This kind of test has not been repeated for more recent versions of the rules, but there is no reason to believe that the quality of speech generated by FOVE would be any poorer. The apparent superiority of synthetic speech is not significant because the variance of the data is so large (only 6 listeners heard the 1974 rule version of the synthetic speech).

It would appear that FOVE is capable of producing speech that is fairly intelligible, especially after a short period of adaptation. There is, however, still room for improvement.

Certainly some of this improvement can be achieved through better specification of the variables available in the FOVE program. Progress will be slow and totally acceptable speech will not be achieved because of limitations of both the hardware synthesizer and the program. A few of these limitations are mentioned below.

Although OVE III produces remarkably good vowels, it does not provide parameters for making nasals and fricatives that closely match real speech. Voiced fricatives are especially poor. There is also the problem of clicks that occur when the periodic (voicing) source is turned off in the middle of a cycle.

Within the program, the major source of frustration for the user is the limited number of environments for which allophone rules can be written. It is simply not possible to specify certain allophonic variants. For others, it becomes necessary to resort to labyrinthine solutions. For example, it may be necessary to specify a minor allophone in the phoneme table and then have a series of rules to change values to the major allophone in a variety of other environments that are provided by the program. Even worse is the situation in which, in order to produce a modification in a limited environment, there must first be a change in a larger context and then one or more rules to change values back to their original specifications in all but the desired context. Not only are such contorted manipulations undesirable on general principles, but because there is a limit on the total number of allophone rules, once the space is used up, no further allophone rules may be written.

Related to the difficulties with the allophone rules is the assignment of distinctive feature numbers that determine which transition values will be used between phonemes. Since the distinctive feature numbers are part of the phoneme table, it is easy to change the number originally assigned to a phoneme by the programmer. However, in so doing, the allophone rule environments also change since the environments have been specified internally in the program by distinctive feature numbers. Thus, advantages achieved in dominance are offset by the loss of certain environments for allophone rules.

There are also problems related to the automatic syllable division of FOVE. These can be overcome by inserting word boundaries wherever necessary and writing allophone rules for word final and word initial allophones, but then any other rules related to word boundary will also apply.

Further problems arise with pitch assignment. Pitch automatically rises on syllables with a major stress, making English sound like a pitch accent language in which pitch rises on accented syllables. Furthermore, there are

problems with final contours because the same pitch specification is used as a component of more than one contour. There is also difficulty in finding single values that are acceptable for both sentences in which a major stress is on the last syllable and sentences in which the last major stress is followed by a number of syllables.

Given a working program that already produces relatively good speech, it might be worthwhile to modify FOVE to give the user access to the component of the program that determines allophone environments. For the production of more natural pitch changes, the user also needs greater control over the fundamental frequency as well. With these kinds of modifications, I believe significant improvements could be made rather quickly in FOVE synthesis by rule.

However, some problems would still remain. The decision was reached at Haskins Laboratories that rather than spend time modifying the existing system, it was time to take what had been learned from FOVE and to design a new program. Accordingly, a new program operating on different principles from its predecessor is now under development (Cooper, Gaitenby, Ingemann, Mattingly, Nye and Shockey, 1977; Cooper, Gaitenby, Ingemann, Levitt, Mattingly, Nye and Shockey, in press). This program is based on user-defined and internal rules that give greater weight to the integrity of the syllable as a unit of speech production. When this new program is available, further work will be necessary to assess its speech quality in relation to FOVE and other synthesis systems.

REFERENCES

Cooper, F. S., J. H. Gaitenby, F. Ingemann, I. G. Mattingly, P. W. Nye and L. Shockey. (1977) Research on audible outputs of reading machines for the blind. Bulletin of Prosthetics Research BPR 10-27, 185-187.

Cooper, F. S., J. H. Gaitenby, F. Ingemann, A. Levitt, I. G. Mattingly, P. W. Nye and L. Shockey. (in press) Audible outputs of reading machines for the blind. Bulletin of Prosthetics Research BPR 10-28.

Holmes, J. N., I. G. Mattingly and J. N. Shearme. (1964) Speech synthesis by rule. Language and Speech 7, 127-143.

Ingemann, F. (1977) Contribution of natural durations to speech synthesized by rule. Journal of the Acoustical Society of America 62, Suppl. 1, S62(A).

Kuhn, G. M. (1973) A two-pass procedure for synthesis by rule. Journal of the Acoustical Society of America 54, 339(A).

Liljencrants, J. C. W. A. (1968) The OVE III Speech Synthesizer. IEEE Transactions on Audio and Electroacoustics AU-16, 137-140.

Mattingly, I. G. (1968) Synthesis by rule of general American English. Supplement to Haskins Laboratories Status Report on Speech Research, SR-33.

Mitchell, P. D. (1974) Test of differentiation of phonemic feature contrasts. Journal of the Acoustical Society of America Suppl. 55, S55(A).

Nye, P. W. and J. H. Gaitenby. (1974) The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences. Haskins Laboratories Status Report on Speech Research SR-37/38, 169-190.

Nye, P. W., F. Ingemann and L. Donald. (1975) Synthetic speech comprehension: A comparison of listener performances with and preferences among differ-

ent speech forms. Haskins Laboratories Status Report on Speech Research SR-41, 117-126.

Shockey, L. (1974) Description of C-MU allophone sentences. ARPA Network Information Center, SUR Note 128.

Segment Duration, Voicing and the Syllable*

Leigh Lisker+

ABSTRACT

The syllable has been proposed as the basic unit of articulatory organization. One kind of evidence cited is the well-known relation between vowel duration and the duration of a following stop; the relatively long voiceless stop is preceded by a shorter vowel than is the shorter voiced stop. Thus the combined durations of vowel and stop "tend" to be equal. One problem with the syllable is that of deciding where to locate its boundary in utterances reported to consist of more than one syllable. The decision criteria most generally used are linguistic (phonotactic) rather than phonetic in nature. Durational data derived from nonsense words of C_1aC_2i type were subjected to the variance test to determine whether durational relations pointed to a syllabification $C_1a \cdot C_2i$ or $C_1aC_2.i$. Aside from the known defects of the variance test, it failed to point consistently to either solution for the different "words" measured. Moreover, the same test failed to yield consistent results when applied to monosyllabic utterances of C_1aC_2.

THE SYLLABLE AS AN INTUITIVE GIVEN

The development of writing has reached full flower in an alphabetic system exemplified by a phonetic transcription in which each so-called speech sound is represented by a unique letter shape. This kind of system has, it is felt, a closer relation in general to the facts of speech perception and production than do other kinds of writing. It is true that representing speech by a sequence of letters, and thus by implication representing it as a sequence of distinct sounds, quite obscures another property of speech signals, namely their quasicontinuous nature. This aspect of speech, however, only becomes apparent when we look at it with the help of laboratory instruments; the ear that listens to speech is still convinced of its discrete (or at least quasidiscrete) nature. (One may wonder whether the lip-reader sees speech as a sequence of discrete elements corresponding to the phonetic segments.) Of course, visible patterns that reflect more faithfully the quasicontinuous nature of speech signals, for example waveforms and spectrograms, seem to be harder to read than any writing system known, and it is possibly true that only a discrete representation can ensure the tolerably

---

*This paper was given at a Symposium on Segment Organization and the Syllable, Boulder, Colorado, 21-23 October 1977.

+Also University of Pennsylvania.

[HASKINS LABORATORIES: Status Report on Speech Research SR-54 (1978)]

efficient transmission of linguistic messages by graphical means. What is somewhat harder to understand is why alphabetic systems fail to incorporate, as a regular feature, any device for indicating the so-called syllabic organization of speech, although the listener is as strongly convinced (we think) that speech is composed of syllables as he is that it is decomposable into sounds. Linguists moreover, particularly phoneticians, show no reluctance about invoking the syllable, and yet the transcriptions they use provide no mark with which to indicate syllable boundaries. If, for example, we are inclined to feel that a word pair like plight-polite differ in syllable structure at least as much as in sound composition (and we might even believe that they differ not at all in the latter), nevertheless our inclination is to diagnose the difference as one of segment composition.

DEFINING THE SYLLABLE: COUNTING VS. DELIMITING

At the same time that the syllable is regarded as a pervasive feature of speech, it is notoriously difficult to make explicit the notion of the syllable or of syllable organization. It has something to do, it seems, with the near-dichotomous classification of the speech sounds into vowels and consonants, and the strong conviction that there exist special relations between a vowel and abutting consonants. These relations are of two kinds at least: in one the vowel and consonant or consonants are intimately connected, and in the other they are products of two distinct articulatory gestures. Speech is then decomposable into subsequences of phonetic elements, each consisting of a vowel and those adjacent consonants having the closer connection to it. Aside from the fact that the separation into vowels and consonants is not always made independently of an analysis into syllables, or at least the identification of syllable "nuclei," it has proved difficult to develop criteria of a purely phonetic nature by which to determine the boundaries between adjacent syllables. Proposals that the syllables be defined language-specifically as an element with respect to which regularities of sound-type distribution may be described have undoubted merit, but the unit so defined is not to be lightly identified with the syllable of our phonetic intuition. The utility of a phonologically defined syllable says nothing one way or the other about whether we may reasonably hope to define the syllable in purely phonetic terms. Of course, if our conviction that it is worthwhile to continue the effort to define the syllable as a phonetic element is based on our phonetic intuition, then we must be clear as to just what that intuition tells us. What it seems to tell us most reliably (how reliably?) is how many syllables comprise an utterance. It informs us much less precisely about where we should feel that one syllable ends and the next begins, if indeed it tells us much at all on this matter that is uncontaminated by knowledge of the phonotactics and grammar of the specific language. The fact that rules devised to decompose speech into syllables are generally not at all what we should call phonetic--they are formulated either on a phonotactic or strictly acoustic basis--would in itself suggest the possibility that syllables are not physically discrete entities at all. Thus, while we are able to locate the syllable "peaks," the "syllabic" elements of the sound sequence, in the case of the consonants we feel impelled to locate them to one side or the other of a syllable boundary; it is not enough for linguists to say simply that they constitute syllable "troughs." Perhaps the problem of locating a syllable boundary is serious only because there is a problem of

spelling encountered anytime we feel the need to hyphenate a word in a phonetically justifiable manner. In introducing a hyphen, we in effect treat syllable boundary as just another element in the linear sequence of letter-shapes. There are, to be sure, two situations in which we are very sure of where to place such a boundary: the onset and the termination of speech coincide, respectively, with the onset and termination of a syllable. It is not obvious, however, that the phonetic properties of speech onset and termination should serve as the basis for analyzing speech into syllables generally.

COARTICULATION: EVIDENCE FOR THE SYLLABLE?

The difficulties encountered in trying to define the syllable as a phonetic unit so that it corresponds exactly to our intuition do not mean that phoneticians refrain from speaking facilely on the subject. As already mentioned, there is the notion that the syllable is a unit of speech production, in that the sequence of sounds composing this unit is generated by a single unanalyzable gesture,[1] and not by a sequence of freely commutable ones. Complementary to this is the assertion that we apprehend speech in units of syllable-size, since there is ample evidence that the phonetic evaluation of an acoustic segment is not context-independent. These phenomena of "coarticulation" (and "coperception") provide some of the strongest motivation for viewing the syllable as a phonetic unit, provided we suppose that these effects are coextensive with the syllable. Coarticulatory effects are all those phenomena that cannot be accommodated by a strictly segmental hypothesis of speech production, according to which the speech sounds of a language are analogous to the letters of typescript, each with a shape unaffected by its neighbors. The evidence that speech is not produced in this way is not necessarily evidence for the syllable, for it is not clear that the scope of all the coarticulatory effects so far observed coincides with the syllable, insofar as its boundaries are well established on intuitive, distributional or other grounds. All that we do know is that when we closely examine the operation of the various structures making up the vocal tract, we find that they do not in general shift position in close synchrony and at the rate at which the phonetic segments are emitted. Thus lip rounding and nasalization, two favorite topics in the study of coarticulation, are segmental features (in English) that refuse to be confined within their "proper" segmental boundaries. It is not obviously true that they pay more attention to syllable boundaries, wherever these may be located. Nasalization, which we associate in English linguistically with consonants, may color a preceding vowel, and rounding, a feature of vowels, may color preceding consonants; we

---

[1]The term "ballistic" has sometimes been applied, but it is not clear what the term conveys. If it can with some plausibility be used in referring to either an opening or a closing movement of the articulators, it seems to stretch the connotation of the word beyond recognition if it is used to describe an opening-closing sequence as a single movement. Moreover, if it is legitimate to call this sequence a single ballistic gesture, why not as well apply the term to a sequence of closing and opening movements?

do not know that such anticipatory coarticulation depends crucially on the absence of a syllable boundary.

Coarticulatory phenomena, which have been understood to tie in somehow with the syllable, are of course measured in time, and the temporal dimension is connected with the syllable in another way also, in that it is the syllable to which we appeal in treating the more narrowly temporal properties of speech, namely tempo and rhythm. Connected with tempo and rhythm, though the connection is a complex one, are the durations of intervals in the speech signal that can be said to correspond more or less to the phonetic segments into which the listener/linguist resolves the signal. An extensive literature (Lehiste, 1970) reporting measured durations of acoustically defined intervals corresponding to vowels and stop-consonant closures has provided evidence that segment duration is determined significantly by a number of factors--by overall speech rate, by articulatory properties of the segment (setting its "intrinsic" duration), by rhythm (stress pattern), and by phonetic properties of neighboring segments. Of these, it is the last that has the most immediate bearing on the question of the syllable; tempo and rhythm clearly affect intervals that comprise more than one syllable, while intrinsic duration is by definition a segmental attribute. The existence of coarticulatory relations between adjacent segments does not constitute evidence for the syllable as a phonetic unit; for that we must find that coarticulatory linkages are markedly weaker between segments said to belong to different syllables.[2]

VOWEL AND STOP DURATIONS IN MONOSYLLABLES

One of the best-known cases of temporal coarticulation is the relation between vowel duration and the voicing of a following consonant, according to which vowels preceding voiceless stops and fricatives are shorter than before their voiced counterparts. This relationship is most clearly seen when we compare isolated monosyllables that differ only in the voicing of their final stops. In a word pair like cop-cob the durations of the vowels, however, defined acoustically, show a ratio of about 2/3. Moreover, if the final stops

---

[2]That differences in linkage between acoustic segments can affect listeners' place of a syllable boundary was demonstrated long ago by Malmberg (1955). Synthetic speech patterns interpreted as vowel-stop-vowel sequences were divided into syllables differently, depending on whether formant transitions were supplied before or after a silent interval corresponding to the stop closure. The stop was grouped with the vowel having the transition. Unfortunately, in natural speech such sequences show transitions both before and after the closure interval, so that, for example, black-out (assuming we agree to put a syllable boundary after /k/) will not lack a transition from the /k/-closure to the following vowel, and the postclosure signal will be a convincing gout when heard alone. If a recording of the expression Greek odds is segmented so that only the postclosure part is heard, the listener will report the same word that he hears when the same operation is performed on the expression Greek gods. The pair can be disambiguated by a talker, but that process certainly involves a disruption of the articulatory "plan" of normal fluent performance.

are released so as to produce a noise that allows us to measure the duration of stop closure from the acoustic record, we find that the duration of /p/ is longer. Since the combined durations of vowel and following closure tend thus to be more nearly equal for the two words than either vowels or stops taken separately, and since vowel and stop must here be tautosyllabic, it seems reasonable to suppose that this temporal interdependence is a feature of syllable organization, especially if it turns out that the entire CVC sequence tends to be of equal duration for the two words. We should then perhaps be able to say that the voicing difference associated with the final stops is to be related very directly to the voicing difference, let us say, between cob and gob, where the onset of voicing relative to the articulatory program of the syllable is differently timed. For cop vs. cob the difference is in the timing of the devoicing gesture relative to the end of the syllable; with the earlier onset of devoicing it happens in English that the accompanying stop closure is also advanced in time--perhaps because articulatory closure is itself a devoicing maneuver! To the extent that temporal relations of this kind can be established for isolated monosyllables, we may then turn to polysyllabic utterances, for example the "words" /kábi/ and /kápi/, whose intervocalic stops are not so readily assigned to one syllable or the other. If we find that the temporal relationship between stressed vowel and following stop is like the one in the monosyllables, then we may claim that the argument for placing a syllable boundary before the stop is thereby weakened. Of course, if no sequence of vowel + stop can be turned up that fails to show this relationship, no matter what its location with respect to the syllable boundary of our phonetic (?) intuition, then the argument for the syllable as a phonetic unit is undermined.

STRESSED VOWEL AND STOP DURATION IN C_1VC_2V

In order to have some specific numbers to talk about in this connection, I had a single native speaker of American English produce, in random sequence, a number (61) of tokens in the following sequences: /kápi/ /kábi/ /kámi/ /gápi/ /gábi/ /gámi/. (The forms with /m/ were included in order to have a bilabial closure associated with nondistinctive voicing.) These were recorded and analyzed spectrographically (Kay Sonograph Model 7029A). The following points were fixed on the spectrograms: onset of the release burst of the initial velar, the onset of voicing following the /k/-aspiration, time of establishment of the bilabial closure and release of the bilabial closure.

CVC as Temporal Unit

Figure 1 shows the frequency distribution of durations measured from onset of the initial burst to the termination of the medial closure, an interval here labeled "stressed syllable," though no prejudice is implied favoring the analysis of CV́CV into CV́C-V over CV́-CV. For this particular speaker, and for the particular set of forms elicited (over six sessions in the course of several weeks), the measured interval averages about 300 msec, with a standard deviation (s) of just over 23 msec. We might want to regard this 300 msec as the "target" or "intrinsic" duration for CV́C-sequences generally, or perhaps for sequences of velar stop + /a/ + bilabial stop. Consideration of the durations of these sequences in the six forms taken separately (Figure 2) suggests that they may differ significantly for the

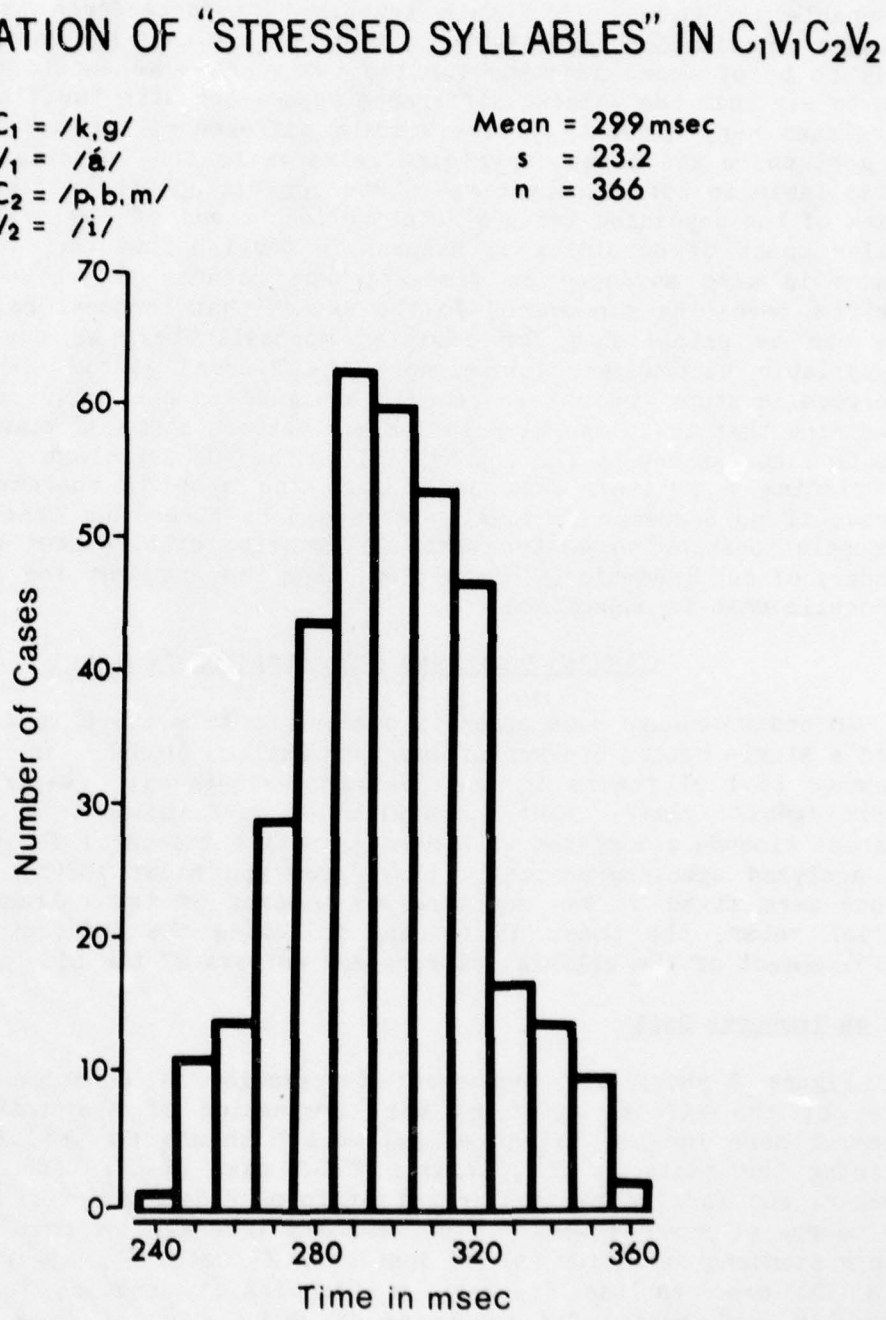


Figure 1: Frequency distribution of durations measured from release of initial velar stop to release of medial bilabial closure. Pooled data of six CVCV sequences.

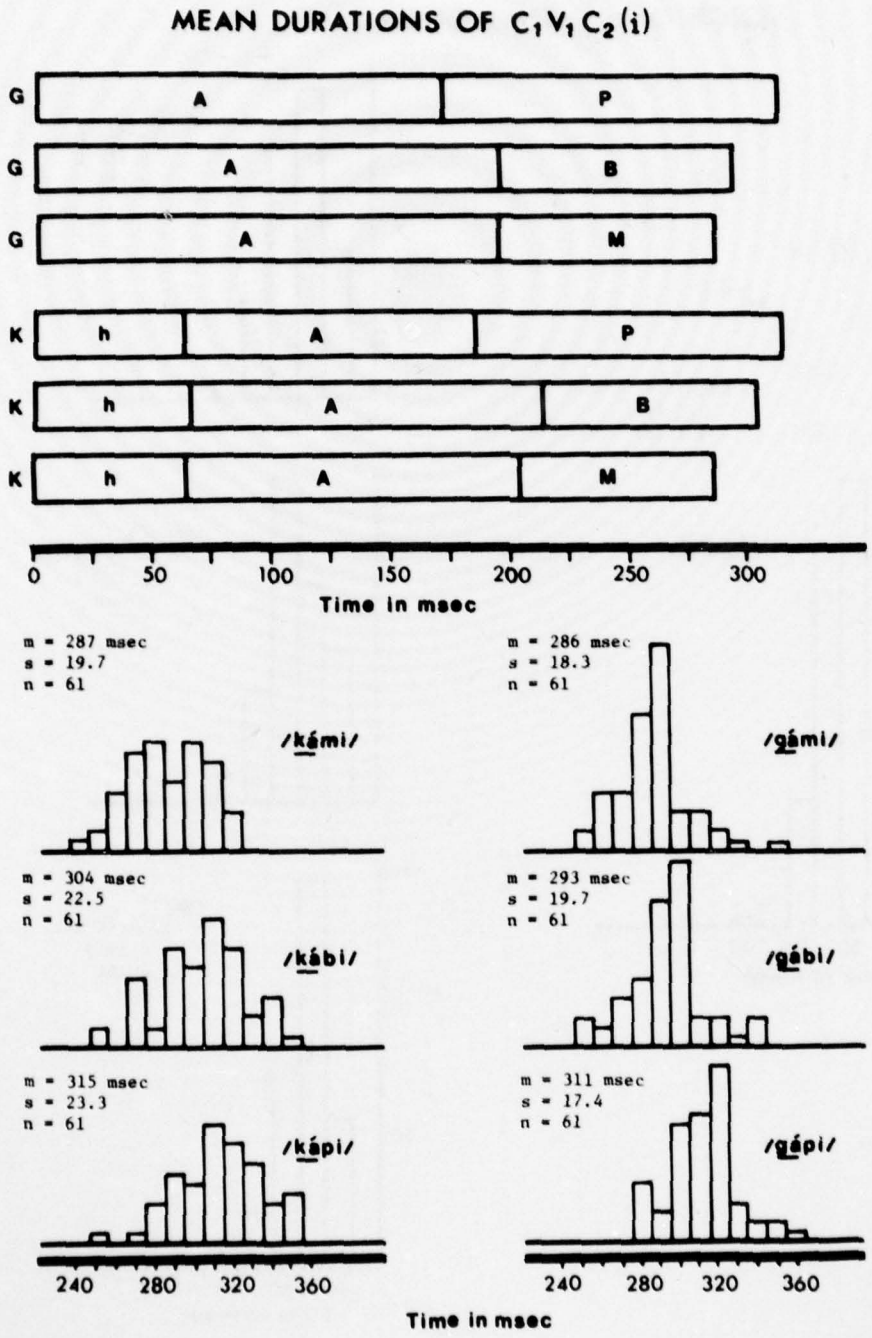


Figure 2: Durations of intervals from initial to medial stop releases. Upper panel: mean durations of measured intervals for each CVCV sequence type. Lower panel: frequency distributions of CVC for the six types.

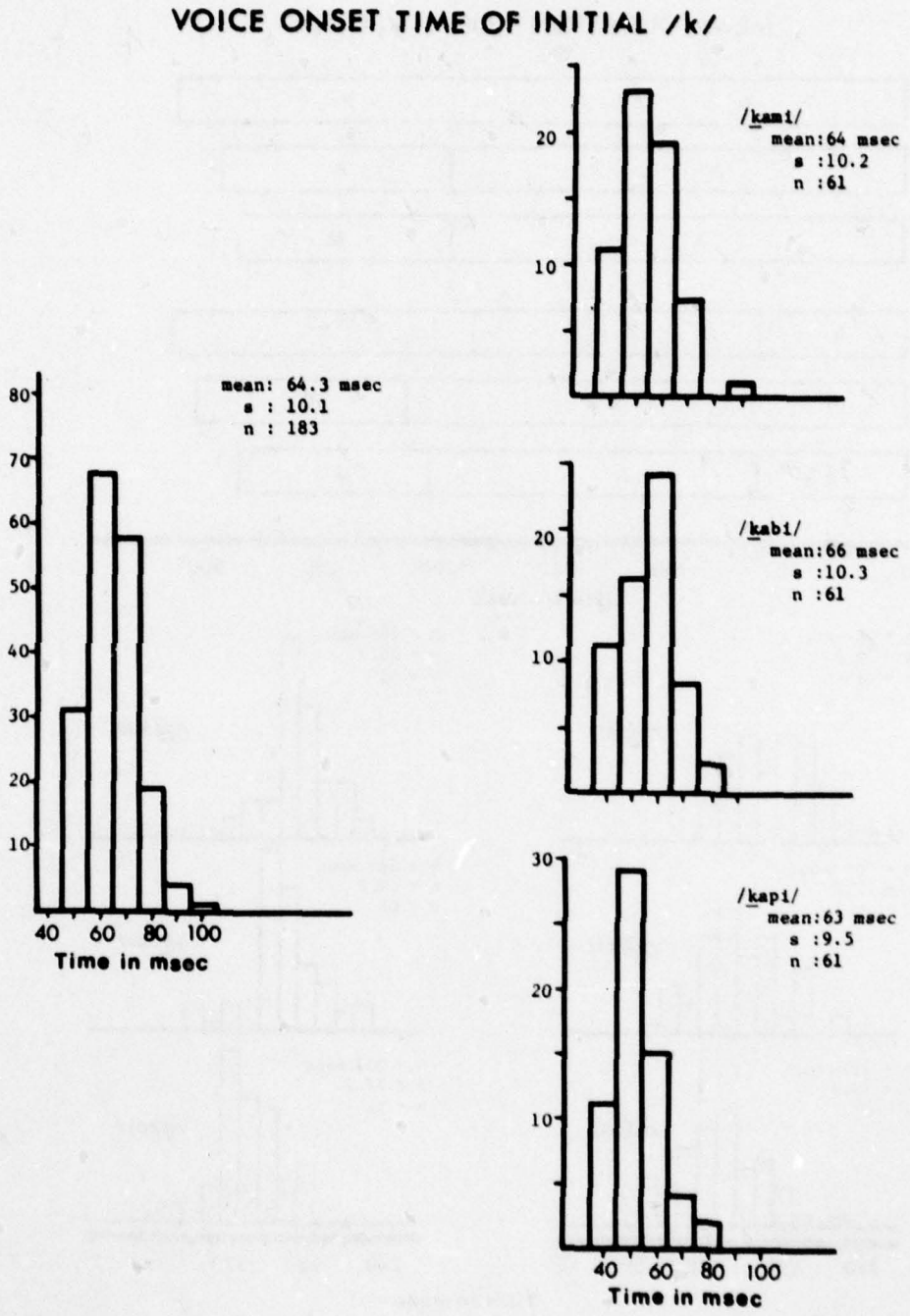
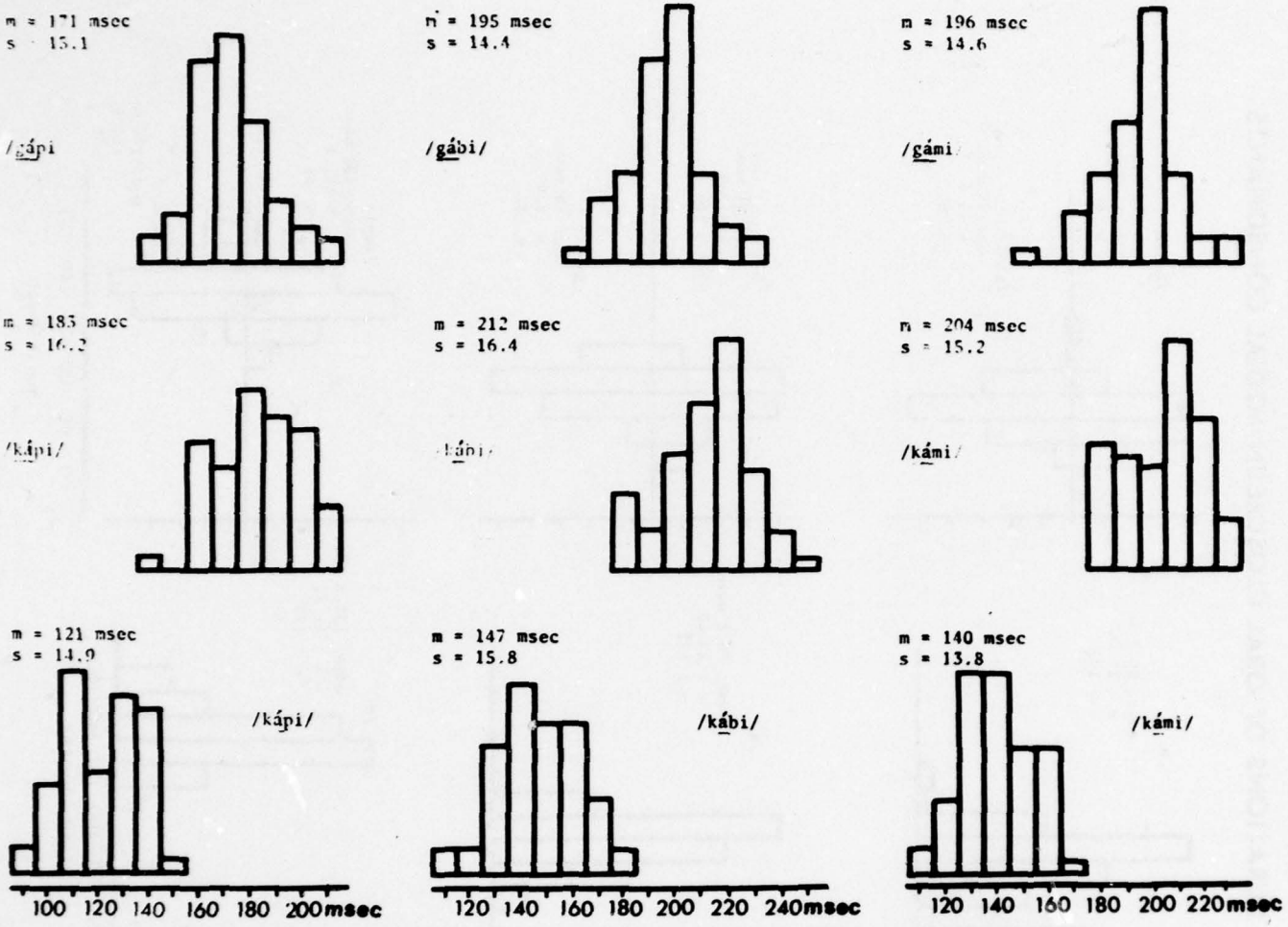


Figure 3: Durations of intervals from /k/-releases to onset of glottal pulsing. Left display represents summation of data shown on right.



DURATIONS OF STRESSED VOWELS BEFORE STOPS IN TROCHEES

Figure 4: Top row: distributions of durations from /g/-release to establishment of medial closure. Middle row: durations of intervals from /k/-release to beginning of medial closure. Bottom row: durations from onset of glottal pulsing following /k/-release to beginning of medial closure.

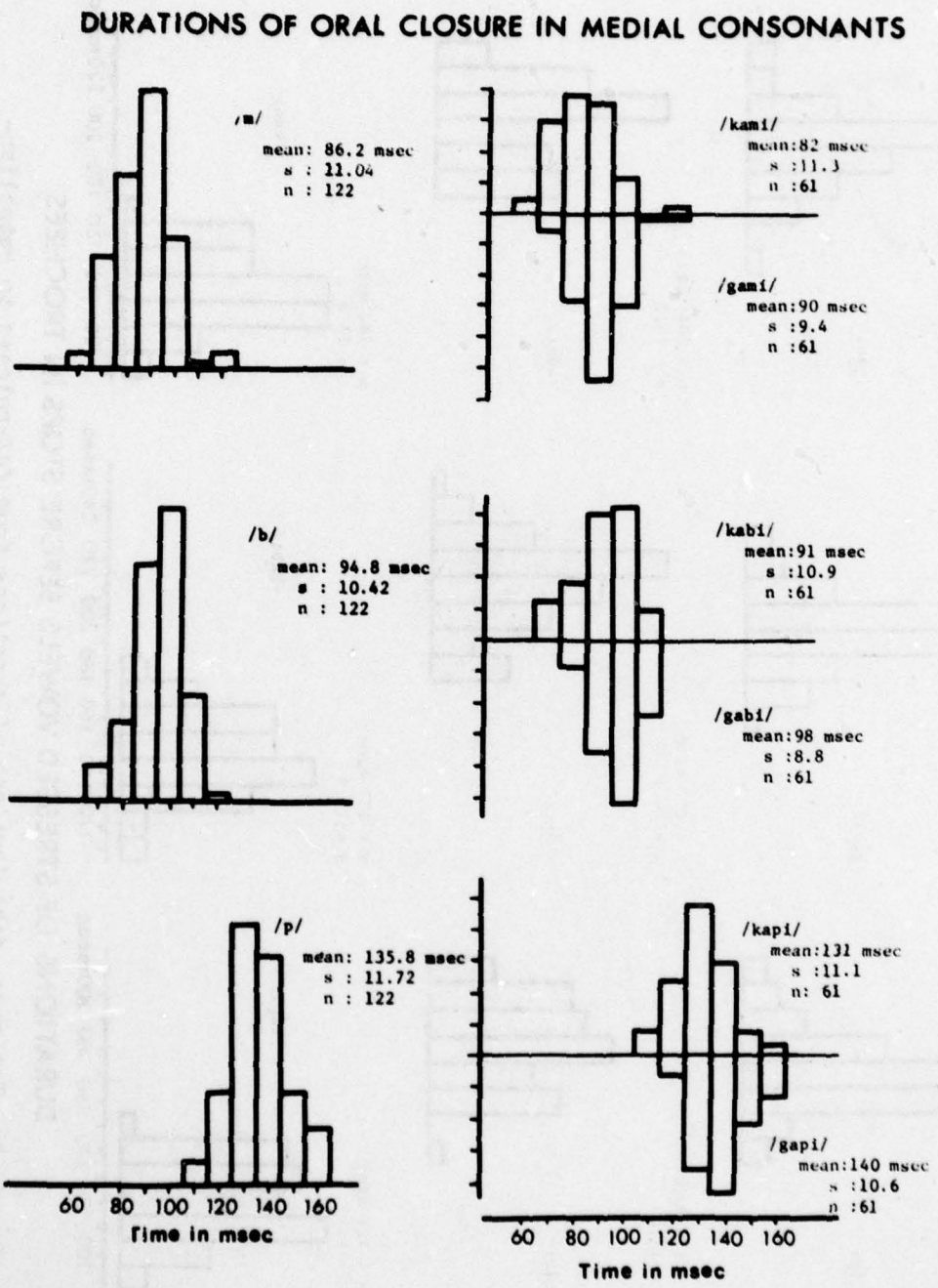


Figure 5: Frequency distributions of medial bilabial closure for /p,b,m/.

different C_1 and C_2. While mean values for /kámi/ and /gámi/ are obviously not different, and /kápi/ and /gápi/ are not clearly different, /kábi/ differs from /gábi/ by an amount that is significant at the .005 level. And even if we entirely discount any differences between /k---/ and /g---/ forms, pairs differing with respect to their medial consonants show mean durational differences that in only one case, /gámi/ vs. /gábi/, fail to meet a criterion of $P < .01$. From Figure 2 it appears that CVC- durations depend significantly on the second consonant: CV́p- > CV́b- > CV́m-. The role played by C_2 is quite evidently more important than any by C_1.

Durations of CV́ and C_2 Taken Separately

Turning now to the component segments to determine how much these vary and to what extent any variability can be ascribed to context, we see, first of all (Figure 3), that /k---/ forms show no significant differences in the intervals between onset of the release burst and onset of laryngeal pulsing. (This interval, the first measured segment of these forms, can be called, as it has been, either "aspiration" or "devoiced vowel," depending on whether one elects to say it is "part of" C_1 or V.) The 3-msec difference between the means of /kábi/ and /kápi/, though there would be no difficulty in devising explanations for it, must be dismissed as statistically nonsignificant.

Instead of deciding what to define as <u>the</u> duration of the stressed vowel (a dubious enterprise of indeterminate meaning), I chose to adopt both practices that have been followed, measuring both from release of the initial stop and from the onset of voicing in the /k---/ forms. In these forms /a/ was taken to correspond both to the entire interval from release to closure (Figure 4, middle row) and to the voiced part of that interval (Figure 4, bottom row). Each of these intervals was compared with the first segments in /g---/, in which any lag in voice onset behind the release was neglected in my measurements. By both measures the /a/-durations differ significantly, depending on both C_1 and C_2; the one exception is in /gámi/ vs. /gábi/. [These relations are very like those reported long ago in Peterson and Lehiste (1960), although their prenasal vowels were longer than those before voiced stops.] The ratios of /á/-durations before /p/ to those before /b/ have a mean value of .85, somewhat greater than the .67 more commonly reported for isolated monosyllables.

The mean values of closure duration are significantly different for the three bilabial consonants (Figure 5): $p < b < m$. Moreover, there seems to be some kind of relation between the voicing of the initial stop and the duration of the medial closure, or perhaps it would be more reasonable to relate the duration of the closure to the duration from initial release to the beginning of the closure interval. The greater duration of closure in the /g---/ forms seems to represent an instance of "compensatory" lengthening, although in the case of /b/ this lengthening was not quite enough to equalize total CVC durations of /kábi/ and /gábi/ (they differ at the .01 level). On the other hand, the significant difference between /m/ and /b/ durations ($P << .01$) is not matched by any appreciable difference in the intervals from /g/-release to onset of medial closure ($P = .05$, i.e., $< .01$). All this says no more than we have already seen, that total durations of CV́C in the CV́CV forms show generally significant ($P < .01$) differences in their means, depending on the

second consonant particularly, and the differences are especially striking when one item of a comparison has the voiceless stop as C_2. (Note, however, that the difference between /kámi/ and /kábi/ is greater than between /kábi/ and /kápi/!).

The Variance Test--Tested and Failed

There is another test that has been applied in similar situations, although its significance is questionable. It compares the sum of the variance of two continuous intervals with the variance of the summed intervals; if the latter quantity exceeds the former, then the two intervals have durations that are negatively correlated. Ohala (1973) has pointed out that variability in locating the boundary between segments will itself yield a negative correlation even if two intervals A and B are truly independent, that is, the relation between their variances (s^2) will be: $s^2(A) + s^2(B) > s^2(A + B)$. On the other hand, a relation $s^2(A) + s^2(B) < s^2(A + B)$ could well result from a variation in "overall speaking rate," a known variable whose effect on segment duration must somehow be cancelled if the "temporal patterning of speech" (a feature that is by definition independent of speech rate) is to be determined. If the first relation can reflect uncertainties in measurements as much as timing "errors" in articulation, and the second results from macroscopic rate variability with no clearly identifiable contribution of the "temporal patterning" factor we are looking for, then a comparison of variances provides data that are part of the problem rather than of the solution. Even if $s^2(A) + s^2(B) = s^2(A + B)$, we are no better off, since there is no reason to exclude this result as a possible outcome of the two competing effects of measurement error and rate variation, given that we may encounter considerable difficulty in making precise determinations of the magnitude of those effects.

Despite these and other caveats that might be raised, the variance relations among the six forms measured are presented in Table 1. If we could ignore the factor of measurement error, we would infer a negative correlation between duration of aspiration and following voiced vocalic interval, while closure intervals are positively correlated with immediately preceding segments, except in the case of /gápi/. These findings are, at least intuitively, in conformity with our belief that the initial stop is tautosyllabic with the following vowel, while the syllable affiliation of the medial stop is uncertain. On the other hand, they are not in conformity with conclusions based on the relations among mean durations of the segments measured, since mean durations of /k/-aspirations show no differences for the different durations of the voiced intervals that follow; in contrast, the clear relationship between the latter interval and closure duration is not apparent in the variance relationship. Since we are not prepared to accept a view that /k/ is not tautosyllabic with /á/, we may by the same token be less willing to use the variance relation between vowel and following closure as the basis for deciding the place of a syllable boundary, that is, analyzing our forms as CV-CV.

On this last point, if the usefulness of the variance relationship has not already been sufficiently impugned by Ohala (1973), we ask what this measure reveals when applied to sequences that are considered monosyllables--

TABLE 1: Variance relations.

s^2	1 (vowel)	2 (closure)	3 (1 + 2)	$\frac{3}{1+2}$
/gápi/	228	112	303	.89
/gábi/	207	77	388	1.36
/gámi/	213	88	335	1.11

s^2	1 (aspiration)	2 (vowel)	3 (1 + 2)	$\frac{3}{1+2}$
/kápi/	90	222	262	.84
/kábi/	106	250	269	.76
/kámi/	104	190	231	.79

s^2	1 (aspiration + closure)	2 (closure)	3 (1 + 2)	$\frac{3}{1+2}$
/kápi/	262	123	543	1.41
/kábi/	269	119	506	1.30
/kámi/	231	128	388	1.08

the English words /kap/ and /kab/. Two sets of ten repetitions of these words were recorded by the same speaker who furnished the dissyllables just discussed. The variance ratios obtained were these:

TABLE 2:

		Trial 1	Trial 2
/kap/:	$\frac{k^h a}{k^h + a}$	.95	.76
	$\frac{ap}{a + p}$	.65	.29
/kab/:	$\frac{k^h a}{k^h + a}$	1.14	1.13
	$\frac{ab}{a + b}$	.73	1.10

If these figures mean anything (and, if anything, how much?), they indicate absence of any clearly negative correlation between aspiration duration and following voiced vocalic interval, possibly a significant difference in the degree of correlation between the duration of the closures of the final stops (all were produced with final releases) and the preceding segments. If we combined the voiceless and voiced intervals between the initial releases and final closures, we obtain the following ratios:

TABLE 3:

		Trial 1	Trial 2
/kap/:	$\frac{k^h ap}{k^h a + p}$	.47	.37
/kab/:	$\frac{k^h ab}{k^h a + b}$	.96	1.21

These figures tempt us to suppose that a negative correlation exists between first and second intervals when /p/ is involved, but not when the final stop is /b/. If the variance ratio tells us something about "syllable organization," it presumably provides no guidance in the matter of syllable boundary placement, since I think we are unprepared to deny monosyllabic status to /kab/ or to believe that /kab/ and /kap/ differ in the number of syllables

composing them.[3] If the differences in variance ratios have any meaning, they indicate only a difference in the temporal organization of the two syllables, nothing about how we should decompose phonetic sequences into syllables. If a ratio greater than unity may hold between segments belonging to the same syllable, in the case of a monosyllable, then the argument for using that ratio as a guide to syllable division becomes suspect. Since the ratios of Table 1 representing relations between medial closures and preceding segments are almost all greater than one, they do not point to any special relation between the two. If these ratios do not give the "right" answer in the case of the monosyllables, why should we use them to resolve the question of syllable division where phonetic intuition speaks with a forked tongue?

REFERENCES

Fujimura, O. and J. B. Lovins. (1977) Syllables as concatenative phonetic units. Contribution to Symposium on Segment Organization and the Syllable, University of Colorado, Boulder Colorado, Oct. 21-23.

Lehiste, I. (1970) Suprasegmentals. (Cambridge: M.I.T. Press).

Malmberg, B. (1955) The phonetic basis for syllable division. Studia Linguistica 9, 80-87.

Ohala, J. (1973) The temporal regulation of speech. Project on Linguistic Analysis, Reports, 2nd Series, no. 17, 1-22.

Peterson, G. E. and I. Lehiste. (1960) Duration of syllable nuclei in English. Journal of the Acoustical Society of America 32, 693-703.

---

[3] In this connection, Fujimura and Lovins (1977) propose that final voiced obstruents be considered "phonetic affixes" rather than "core" constituents of syllables. The data presented here are consistent with their proposal. However, an analysis of spoken English into sequences of syllabic "cores" and "affixes" would yield elements rather different from the syllables of our phonetic intuition.

II. PUBLICATIONS AND REPORTS

III. APPENDIX

PUBLICATIONS AND REPORTS

Fowler, C. A., I. Y. Liberman and D. Shankweiler. (1977) On interpreting the error pattern in beginning reading. Language and Speech 20, no. 3, 162-173.

Freeman, F. J., E. S. Sands and K. S. Harris. (in press) Progressive changes in articulatory patterns in verbal apraxia: A longitudinal case study. Brain and Language.

Harris, K. S. (1978) Physiological aspects of speech production. In Speech and Language in the Laboratory, School and Clinic, ed. by James F. Kavanaugh and Winifred Strange. (Cambridge: M.I.T. Press).

Lisker, L., A. M. Liberman, D. M. Erickson, D. Dechovitz and R. Mandler. (1977) On pushing the voice-onset-time (VOT) boundary about. Language and Speech 20, no. 3, 209-216.

Mermelstein, P. (1978) Difference limens for formant frequencies of steady-state and consonant-bound vowels. Journal of the Acoustical Society of America 63, no. 2, 572-580.

Mermelstein, P. (1978) On the relationship between vowel and consonant identification when cued by the same acoustic information. Perception & Psychophysics 23(4), 331-336.

Repp, B. H. (1978) A note on single- and double-correct responses in the dichotic two-response paradigm. Journal of the Acoustical Society of America 63, 1220-1222.

Repp, B. H., A. M. Liberman, T. Eccardt and D. Pesetsky. (in press) Perceptual integration of temporal cues for stop, fricative, and affricate manner. Journal of Experimental Psychology: Human Perception and Performance.

Repp, B. H. (1978) Stimulus dominance and ear dominance in the perception of dichotic voicing contrasts. Brain and Language 5, 310-330.

Sands, E. S., F. J. Freeman and K. S. Harris. (in press) Progressive changes in articulatory patterns in verbal apraxia. Brain and Language.

APPENDIX

DDC (Defense Documentation Center) and ERIC (Educational Resources Information Center) numbers SR-21/22 to SR-54:

Status Report		DDC	ERIC
SR-21/22	January - June 1970	AD 719382	ED-044-679
SR-23	July - September 1970	AD 723586	ED-052-654
SR-24	October - December 1970	AD 727616	ED-052-653
SR-25/26	January - June 1971	AD 730013	ED-056-560
SR-27	July - September 1971	AD 749339	ED-071-533
SR-28	October - December 1971	AD 742140	ED-061-837
SR-29/30	January - June 1972	AD 750001	ED-071-484
SR-31/32	July - December 1972	AD 757954	ED-077-285
SR-33	January - March 1973	AD 762373	ED-081-263
SR-34	April - June 1973	AD 766178	ED-081-295
SR-35/36	July - December 1973	AD 774799	ED-094-444
SR-37/38	January - June 1974	AD 783548	ED-094-445
SR-39/40	July - December 1974	AD A007342	ED-102-633
SR-41	January - March 1975	AD A103325	ED-109-722
SR-42/43	April - September 1975	AD A018369	ED-117-770
SR-44	October - December 1975	AD A023059	ED-119-273
SR-45/46	January - June 1976	AD A026196	ED-123-678
SR-47	July - September 1976	AD A031789	ED-128-870
SR-48	October - December 1976	AD A036735	ED-135-028
SR-49	January - March 1977	AD A041460	ED-141-864
SR-50	April - June 1977	AD A044820	ED-144-138
SR-51/52	July - December 1977	AD A049215	ED-147-892
SR-53	January - March 1978	AD A055853	**
SR-54	April - June 1978	**	**

AD numbers may be ordered from:

U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Road
Springfield, Virginia 22151

ED numbers may be ordered from:

ERIC Document Reproduction Service
Computer Microfilm International Corp. (CMIC)
P.O. Box 190
Arlington, Virginia 22210

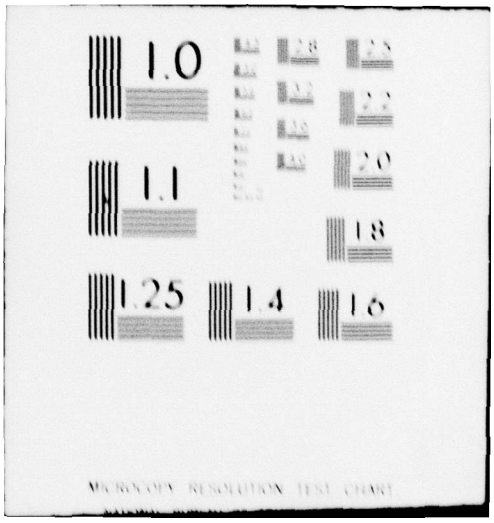
**DDC and/or ERIC order numbers not yet assigned.

Haskins Laboratories Status Report on Speech Research is abstracted in Language and Behavior Abstracts, P.O. Box 22206, San Diego, California 92122.

AD-A060 448 HASKINS LABS INC NEW HAVEN CONN F/G 5/7
SPEECH RESEARCH. A REPORT ON THE STATUS AND PROGRESS OF STUDIES--ETC(U)
JUN 78 A M LIBERMAN V101(134)P-342
UNCLASSIFIED SR-54(1978) NL

3 OF 3
AD A060 448

END
DATE FILMED
1-79
DDC


1.0
1.1
1.25
1.4
1.6
2.0
1.8
MICROCOPY RESOLUTION TEST CHART

Errata

SR-53, vol. 1

Fowler and Turvey. Skill Acquisition: An Event Approach with Special Reference to Searching for the Optimum of a Function of Several Variables.

p. 158, Para. 2, l. 1	change E to ΔE
p. 158, Para. 3, l. 2	change E to ΔE
p. 159, l. 3	change (ΔE) to $\Delta(\Delta E)$
p. 160, l. 7	change (ΔE) to $\Delta(\Delta E)$
p. 160, Table 3, l. 3	change (ΔE) to $\Delta(\Delta E)$

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
Haskins Laboratories 270 Crown Street New Haven, Connecticut 06510	UNCLASSIFIED
	2b. GROUP N/A

3. REPORT TITLE

Haskins Laboratories Status Report on Speech Research, No. 54, April-June, 1978

4. DESCRIPTIVE NOTES (Type of report and, inclusive dates)

Interim Scientific Report

5. AUTHOR(S) (First name, middle initial, last name)

Staff of Haskins Laboratories; Alvin M. Liberman, P.I.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
June 1978	199	279

8a. CONTRACT OR GRANT NO.		9a. ORIGINATOR'S REPORT NUMBER(S)
HD-01994	NS13870	SR-54 (1978)
V101(134)P-342	NS13617	
N01-HD-1-2420		
RR-5596		9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
BNS76-82023		None
MCS76-81034		

10. DISTRIBUTION STATEMENT

Distribution of this document is unlimited*

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
N/A	See No. 8

13. ABSTRACT

This report (1 April - 30 June) is one of a regular series on the status and progress of studies on the nature of speech, instrumentation for its investigation, and practical applications. Manuscripts cover the following topics:

-Categories and context in the perception of isolated steady-state vowels;
-Tongue position in rounded and unrounded front vowel pairs;
-The reading behavior of dyslexics: Is there a distinctive pattern?;
-Articulatory units: Segments or syllables;
-Selective anchoring and adaptation of phonetic and nonphonetic continua;
-Speech across a linguistic boundary: Category naming and phonetic description;
-Discrimination of subphonemic phonetic distinctions;
-Anticipatory coarticulation: Some implications from a study of lip rounding;
-Rapid vs. rabid: A catalogue of acoustic features that may cue the distinction;
-Acoustic characteristics of normal and pathological voices;
-Speech synthesis by rule using the FOVE program; and
-Segment duration, voicing and the syllable.

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6811

UNCLASSIFIED

Security Classification

A-31408

*This document contains no information not freely available to the general public. It is distributed primarily for library use.

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Vowels - categories, contexts						
Tongue - articulation, vowels						
Reading behavior - dyslexia						
Articulation - segments, syllables						
Adaptation and anchoring - phonetic, nonphonetic						
Linguistic behavior - boundary, category, phonetic						
Acoustic features - speech cues						
Duration - segment, voicing, syllables						
Speech discrimination - phonetic						
Acoustic features - vocal pathology, normal, abnormal						
Coarticulation - anticipation, lip rounding						
Speech synthesis - computers, phonetic segments, rules						